Guarding the Guardians

Rating scale and rater training effects on reliability and validity of scores of an oral test of Norwegian as a second language.

Cecilie Carlsen

2003

Nordisk institutt
Universitetet i Bergen
Bergen, Norge

Table of contents

Acknowledgements

Writing this thesis would have been a lot harder if it hadn't been for the help and support of a number of wonderful people. First of all, I must thank my supervisor, Jon Erik Hagen, for his big heart and clear mind, for fruitful discussions and critical comments, but most of all for his constant faith in me. Benedicte, for being my sister and dear friend since the dawn of my memory, for reading through the thesis, and for making me see its funny sides.Tania Horak, for proof-reading on such a short notice! Colleagues and friends at Norsk språktest, for showing me professional language testing in practice, as well as for support, chocolate and memorable study-trips abroad. My friends, for sharing everything. My parents, for shelter in stormy weather. My sons, Pål and Tarje, for teaching me the difference between important and less important things.
Jørgen, for making life so sweet.

Abstract

This thesis focuses on the scoring of a national test of Norwegian as a second language: *Språkprøven i norsk for voksne innvandrere,* developed by Norsk språktest at the University of Bergen. In order to ensure a fair assessment of the candidates' oral production, the test constructors make use of *trained raters* basing their scores on an *explicit rating scale (NORS)*. These two highly recommended procedures in performance testing have traditionally been viewed as means to heighten reliability of test scores. In line with recent developments in the field of language testing, I argue that the rater variable affects not only reliability, but the very construct validity of test scores. Rater training and development of rating scales are costly and time-consuming enterprises. To establish their effect on test scores is therefore interesting from a test theoretical, as well as from a practical and economical point of view.

In the study, four groups of informants are compared: non-linguists (or naïve-native speakers), teachers of Norwegian as a second language without rater training, raters of Språkprøven, and finally a subgroup of the most experienced raters of Språkprøven. The informants score eight candidates' video recorded performances on a six-point scale. The first four are scored impressionistically, and the next for by informants using the NORS. The quantitative data are used in an investigation of internal agreement (inter-rater reliability) between raters of the distinct groups. Informants are also asked to give written reports of their scores, which are used in an investigation of raters' underlying criteria for assessing speech. The qualitative data are used firstly in an attempt to explain the results of the reliability study, and thereafter in an investigation of the match between raters' criteria and the criteria of the NORS (construct validity). The results reveal differences between groups for the scores they give, as well as for the reasons for these scores. One important conclusion echoes the claim that "quantitative similarities in ratings may mask significant qualitative differences in the reasons for those ratings" (Connor-Linton 1995: 99).

Abbreviations

AFL	Arabic as a foreign language
ALTE	The Association of Language Testers in Europe
ACTFL	American Council on the Teaching of Foreign Languages
CC	communicative competence
CEF	Common European Framework of Reference for Languages
CLA	Communicative language ability
CR	criterion-referenced
CTT	classical test theory
EFL	English as a foreign language
ESL	English as a second language
FSI	Foreign Service Institute
GB	government and binding
H	hypothesis
ILR	Interagency Language Roundtable
IRR	inter-rater reliability
IRT	item-response theory
KAL	Kvalitetssikring av læringsutbyttet i norsk skriftlig
L1	first language
L2	second language
LT	language testing
MFR	multi-faceted Rasch analysis
MM	multidimensional model
MTMM	multi-trait multi-method
MTS	multiple trait scoring
N2	Norwegian as a second language
NNS	non-native speakers
NORS	Norwegian oral rating scale
NR	norm-referenced
NS	native speakers
NST	Norsk språktest
PT	processability theory
RP	received pronunciation
RQ	research question
SL	second language
SLA	second language acquisition
SP	Språkprøven
UG	universal grammar
WR	written reports
ZISA	Zweitspracherwerb Italienischer und Spanischer Arbeiter.

List of Tables

List of Figures.

A study of [the history of language testing] reveals the continuing tension between the demands of psychometric theory and practice for objectivity and reliability in measurement, and the fact that what is being measured is that most flexible, multidimensional, fugitive, and complex of human abilities, the ability to use language
(Spolsky 1995:39).

CHAPTER 1 INTRODUCTION

The focus of this project is on the measurement and scoring of speech in a second language (L2) context in relation to a national test of Norwegian as a second language (N2), developed by Norsk språktest at the University of Bergen. In the field of language testing (LT) difficulties in relation to oral testing have long been recognised (Spolsky 1995, Bachman 1981, Fulcher 1997, McNamara 1995). Bachman claims that: "One of the areas of most persistent difficulty in language testing continues to be the measurement of oral proficiency" (Bachman 1981:67). The difficulties connected to the measurement of speech also apply to research into the qualities of oral tests, as pointed out by Fulcher:

> The criteria by which a good test of speaking can be judged are those which can be applied to all language tests: reliability, validity and practicality. In the testing of speaking, however, the problems in examining these qualities are heightened by the nature of speech itself. Eliciting a large enough language sample for adequate assessment is time consuming and expensive, while scoring will for the foreseeable future depend on the use of expert human judges (Fulcher 1997:75).

Firstly, in what ways does the spoken language "by nature" make testing and test-research more difficult than for example the skill of writing? One obvious reason is that while a written text exists for as long as someone takes care of the sheet of paper on which it is written, a spoken text only exists for as long as the sound waves hang in the air (unless it is tape- or video-recorded, of course). When studying written texts we can read a passage, go back and re-read parts of it, we can take a break if we are tired etc. When faced with an orally delivered text, on the other hand, interlocutors, raters and researchers have to pay full attention for as long as it lasts. It is not always possible or appropriate to ask for clarifications or repetitions. When studying the written language, researchers may in many cases retrieve data from paper archives or electronic corpora. Such corpora are less common for the oral language, and in most cases researchers have to gather their data from time- and cost-consuming face-to-face contact with their informants. This, of course, has consequences for the size of the data set in the treatment of speech.

However, problems related to the measurement and study of speech are also due to the lack of good theories and models describing it (Bygate 1987: *vii*, Fulcher 1997:82, Saleva 1997: 18*)*. We still know more about the rules governing writing than about the rules governing speech. Discourse analysis and interactional analysis are important contributions to the understanding and description of speech, but these approaches have only to a very small extent been used as the basis for tests. Rather, oral test construction after 1990 has almost exclusively built upon general

models of communicative language ability as presented in Bachman (1990) and Bachman and Palmer (1996). These models do not, however, distinguish between the writing and speaking skills, nor do they describe the characteristics of each mode. The lack of a complete model for the oral mode may be one reason for the seeming fact that speech is ephemeral and hard to get hold of.

The other cause of difficulty mentioned by Fulcher is the way that oral tests are scored. As opposed to the typical multiple-choice tests, which may be scored according to a key by a lay-person or even a computer, tests of oral or written production cannot be scored without the evaluation of human raters (McNamara 1996:117). For these skills, it is not a question of true or false as much as a question of good or poor, hence a question of quality. And as in other areas of life, quality is a matter of taste; it is a matter of subjectivity. The subjective aspect of the rating process represents an important, if not *the* most important, challenge of oral testing. How can we trust raters to focus exclusively on the oral performance of the candidate and not on irrelevant aspects such as the candidate's sex, nationality, personality, general knowledge of the world and apparent intelligence? And even if raters manage to keep these other aspects apart and focus on the speaking skill alone, how can we be sure that they evaluate speech in the same way? Moreover, can we take it for granted that candidates are rated identically by different raters? Or that one rater gives two candidates who perform alike the same score? Professional language testing normally meets these challenges by standardising the scoring procedure. This includes three different procedures. Firstly, test constructors define the construct of the test and develop *rating scales* which operationalise it. The rating-scale defines typical performance at different levels on the numeric grade-scale according to a set of rating criteria. Secondly, the test constructors *train their raters* in how to interpret and use the rating-scale in their evaluation in order to make raters score more self-consistently on the one hand and more in accordance with each other on the other. An additional value of the training sessions is to ensure that raters focus on the traits which constitute the construct of the test and not on other irrelevant aspects of performance and thereby jeopardise the construct validity of the scores. Yet, it is important to emphasise that even though the use of rating scales and training of raters may increase the validity and reliability of test scores, human evaluation of this kind may never be made totally objective. As McNamara states:

> The assumption in most rating schemes is that if the rating category labels are clear and explicit, and the rater is trained carefully to interpret them in accordance with the intentions of the test designers, and concentrates while doing the rating, then the rating

> process can be made objective. […] But the reality is that rating remains intractably subjective (McNamara 2000:37).

Facing the consequences of this point, a third procedure is necessary: The candidates should be evaluated by *multiple raters* whose scores should be added and averaged to reach the most reliable assessment possible (Alderson 1991b:68). The more raters, the more reliable the scores Yet, for economic reasons, more than two raters are seldom used.

The development and use of explicit rating scales and the training of raters are procedures that are taken to enhance reliability and validity of scores. Despite the extensive use of these procedures, we still do not know very much about their effects on test scores (McNamara 1996:126, Weigle 1994:7).

For a test based on the subjective evaluation of human raters to yield fair scores, it is of major importance that the raters perform their task in a satisfactory way. If not, the scores will be unpredictable, the test takers will not be given a fair judgement, and society cannot rely upon the scores of the test. It is therefore important to keep an eye on the raters, to guard the guardians, to make sure their ratings are up to standard. The purpose of the thesis is not, however, limited to the evaluation of one group of raters of one particular test of Norwegian as a second language (N2), which would be of limited interest.

1.1 Research questions and hypotheses

The overall aim of the project is to investigate the rater effect on test scores. In an effort to reach this aim, I raise several research questions and a series of hypotheses are formulated. The research questions relate to the effect of rating scales and rater training on the agreement between raters about scores (inter-rater reliability, IRR), and the agreement between the criteria raters use and those of the rating scale (construct validity). They are as follows:

> RQ1: Does the use of trained raters and a rating scale produce raters who are more in agreement about the scores they give, that is, do these procedures have a positive effect on inter-rater reliability?
>
> RQ2: Does the use of trained raters and a rating scale produce raters who are more in agreement with the test constructors about the underlying construct of the tests as specified in the rating scale, in other words, do these procedures have a positive effect on construct validity?

From these principal research questions, four hypotheses are deduced, two of which regard the effect of these procedures on IRR, and two regarding their effect on construct validity:

- H1: **Training of raters** affects **reliability** of scores positively; trained raters show higher inter-rater reliability than untrained raters when scoring both with and without rating scales.

- H2: The use of an explicit **rating scale (NORS)** affects **reliability** of scores positively; inter-rater reliability of scores is higher when raters use a rating scale (the NORS) as opposed to when they score impressionistically. The effect of a rating scale is positive for raters with and without rater training, yet the effect is greatest for the groups of untrained raters (naïve NS and N2-teachers).

- H3: **Training of raters** affects **construct validity** (defined as the match between the criteria of the scale and those of the raters) positively: there is a greater match between the criteria of the NORS and those of the trained raters than between the NORS and the criteria used by other rater groups.

- H4: The use of an explicit **rating scale (**NORS) affects **construct validity** (as defined in H3) positively. There is a greater match between the criteria of the NORS and those of the raters when raters base their scores on the NORS than when scoring impressionistically.

1.2 Data and methods

The project has been conducted using a hypothetical-deductive approach. I have attempted to falsify the four hypotheses against different kinds of empirical data. The study of IRR (H1 and H2) has been treated quantitatively, while the questions related to raters' interpretation and use of the criteria (H3 and H4), have been investigated through a qualitative method for the most part. The combination of quantitative and qualitative approaches has been done in an effort to grasp more of the whole picture of rater-behaviour and rater-effects than would be possible by using only one of the approaches in isolation. A combination of the two kinds of data is also recommended by other researchers in the field (Bachman 1997, Weigle 1994, Tarnanen 2002)

The data of the study can be divided in three:

- background-information about the raters
- raters' numeric scores
- raters' written reports(WR) for the scores they awarded

The background information about the raters has been used mainly as a basis for a categorisation of raters into four groups: "naïve NS", "teachers of Norwegian as a second language", "raters of Språkprøven" and finally "expert-raters of Språkprøven".

Raters' numeric scores have been used in the investigation of the possible effect of rater training and rating scale on IRR of test-scores.

Raters also gave written explanations for each of the scores they gave. These written reports constitute the qualitative data of the study. There was no guidance whatsoever as to the explanations they could give, but they were asked to argue both why they did not give a higher score and why they did not give a lower score. (Halleck 1992). In order to facilitate the analysis of these data, the criteria have been categorised and coded for statistical analysis.

There are three more potential data sources of the project, which have been treated only superficially. The first is the rating scale, the NORS, itself. Obviously, raters have to interpret the descriptors of the rating scale in order to apply it. A poorly defined rating scale may therefore affect the way in which it is applied. Moreover, for a test to yield valid scores, the rating scale needs to be a valid operationalisation of the construct of the test. These questions in relation to the quality and validity of the rating scale are only handled briefly in this thesis. Its aims did not include establishing the construct validity of any test in particular. Whether or not the scale is valid is therefore considered subordinate, or in other words, I take as a premise for my investigation that the NORS is a valid representation of the construct of Språkprøven.

The second data-source not treated in detail here is the oral performance of the eight non-native speakers (NNS). It would have been interesting to investigate the match between raters' comments with a thorough analysis of the candidates' performances. This question has not been followed up in this thesis, but it is handled in detail in a follow-up study[1].

The third kind of data not exploited in this thesis, is whether raters use a certain criterion positively or negatively in their written reports. It is possible that some traits are used in raters' argumentation for why they did not give a higher score, while others are used in their reasons why they did not give a lower score. These differences in the use of criteria are also investigated in the follow-up study mentioned above.

1.3 The scientific value and generalizability of the study

Why is a study like the present one scientifically relevant and interesting? In what ways does it contribute to the LT-field with new information? I would claim that its unique value lies at three levels: its overall research focus, the complexity of its design and the language under study.

[1]Carlsen (In progress): "Lekfolk og fagfolks vurdering av aksentpreget norsk".

The first value of the project lies in the *research focus*. The project highlights the effect on reliability and validity of two highly recommended and commonly used procedures in performance-testing: the training of raters and the use of explicit rating scales. The effect is investigated in relation to inter-rater reliability as well as to the construct validity of test scores. To establish the effect of these procedures is of theoretical as well as practical and economic concern: in the field of language testing, the rater variable has traditionally been taken to affect the reliability of test scores. In modern test theory, however, one is starting to realise that the rater variable affects the very construct validity of test scores. If raters fail to focus on the construct operationalised in the rating scale, this is assumed to affect the validity of scores. The test will no longer measure what it sets out to measure. This study is one contribution to that discussion of the role of the rater variable on test scores.

The use of rater training and the development of rating scales are time- and cost-consuming procedures. It is therefore of great practical and economic interest to establish whether or not they have the intended effect on test-scores, and hence whether they are worthwhile.

There have been some studies focusing on the effect of rater-training on test-scores (Shohamy et al 1992, Sieloff Magnan 1987, Cumming 1990, Vaughan 1991) but results are inconclusive as to whether or not trained and experienced raters are more suitable raters than people without rater-training. All of the studies mentioned above compare groups of experienced and inexperienced raters in their investigations. Weigle's study of the effect of rater-training is interesting because she uses a before-after design (Weigle 1994, 1998). The drawback of this design is that it does not capture differences between groups due to their degree of rater experience. It is possible that rater-training has a positive, yet not immediate effect. Lumley and McNamara (1995) explore whether the effect of rater training is stable over time. My study takes into account the effect of rater training by comparing two groups of teachers of Norwegian as a second language (N2-teachers) of which one group has received rater-training and the other has not. In addition, the group of experienced raters is further subcategorised according to their varying degree of experience as raters, and they are compared with a group of linguistically naïve native speakers, functioning as a control group. In my search in the LT-literature, I have not managed to find examples of studies comparing scores given by the same groups of raters when scoring impressionistically on the one hand and with an explicit rating-scale, on the other. However, Shohamy et al (1992: 31) do call for this kind of research. In the present study, the four rater-groups use both scoring-methods: they first evaluate four candidates without any kind of guidance, totally subjectively and impressionistically. In the assessment of the next four

candidates they base their scores on the NORS, which highlights the criteria by which the candidates are to be tested and exemplifies performance at different levels. This is, as far as I am aware, innovative.

In the discussion in Chapter 11, I investigate possible reasons for the results of the reliability study in the criteria raters use. I discuss the assumptions that the use of few criteria over many, the focus on formal linguistic traits over traits related to communicative ability, and internal agreement about the underlying criteria enhance inter-rater reliability. This linking of rater reliability and raters' use of criteria has, to my knowledge, not been investigated in earlier studies.

The *complexity of the design* is another value of the study: I combine two scoring methods, impressionistic and scale- based, four rater-groups (naïve NS, N2-teachers, raters of Språkprøven with rater training but limited experience, and raters of Språkprøven with training and extensive rater experience) and two data-types (numeric scores and written reports). This gives a complex design combining quantitative and qualitative data, allowing many interesting research questions to be raised. I have touched upon some of them in this thesis, but the data allow many more interesting research questions to be investigated. I consider the collection of a rather large set of data for further research an important feature of the project.

A third value of this study lies in the fact that the *language in focus* is Norwegian. Most of the research-literature on the rater-variable is related to English. Often the research questions generalise to the testing of other languages as well, but it is important that research is conducted on other languages too, a point also made by Chalhoub-Deville (1995: 28).

Finally, the project is valuable as a contribution to the field of language testing which is in its very infancy in Norway. During the last couple of years, there has been a few research projects focusing on test related issues (Berge 1996, Hasselgren 1998, the KAL-project). Despite these pioneer studies, more work needs to be done in order to establish and develop the field of language testing in Norway which is necessary in order to develop fair and reliable measurement instruments aimed at different levels of the education system as well as for adults and immigrants. This project is one contribution to this field of research.

1.4 Organisation of the thesis

This thesis is organised into two main parts, one theoretical and one empirical. The first two chapters present a linguistic framework, while the others handle theoretical issues in relation to the measurement of speech. In the second half the empirical investigation of the study is treated.

The study belongs to the field of language testing and has reaped its fruits from two different, though related theoretic orchards: general linguistic theories on one side of the fence and test-theory and psychometrics on the other. For a language test to be valid a good description of its theoretical construct is of fundamental importance. Oral language testing in the communicative approach of the 80s and 90s has its basis almost without exception in a general model of communicative language ability as presented by Bachman (1990:87). The historical evolution of the concept of CC as well as different models describing it are presented in Chapter 2. Despite the central position of models of CC as basis for language test construction, it has been argued that they are not suited as thorough descriptions of the spoken mode. Saleva argues that "[…] the most commonly used versions of communicative competence do not make any distinction between oral and written proficiency […]" (Saleva 1997:13). A discussion of what oral proficiency is seems necessary as a starting point for an investigation into a test of this ability. This is the focus of Chapter 3.

Chapter 4 introduces a shift in focus from linguistic theory to issues in test theory. It serves the purpose of a general introduction to language-testing and some key-concepts of the field are presented and defined. The next three chapters focus on distinct aspects of the assessment of speech. Different approaches to testing speech are discussed in Chapter 5, rating scales and rating criteria for speech are handled in Chapter 6, and finally the rater-variable is the focus of Chapter 7.

Chapter 8 introduces the second main part of the thesis, the one covering the empirical investigation. The method and design of the investigation are presented in Chapter 8. As the design is rather complex, the results are presented in two chapters. Chapter 9 presents the results of the reliability study, which is based on the quantitative data, and in Chapter 10 the results of the validity study and qualitative data are presented. Finally, in Chapter 11, the results of the study are discussed and the qualitative data are used in a tentative explanation of the results of the reliability study. Chapter 12 presents a summary of the results, and some conclusive remarks about the investigation are outlined. The thesis is rounded off with some theoretical and practical implications of the study, as well as some suggestions for future research projects on rater related issues.

CHAPTER 2. COMMUNICATIVE COMPETENCE AS A BASIS FOR LANGUAGE TESTING

"[…A] language test is only as good as the theory of language on which it is based" (McNamara 2000:86). McNamara's statement highlights the close connection between language testing and linguistic theory. A language test will always rest on a certain conception of what language is. In the field of LT the importance of basing language tests on linguistic theory was emphasised by Lado as early as the beginning of the 1960s in what is often considered the first book dedicated specifically to the emerging field of language testing: *Language Testing: the Construction and Use of Foreign Language Tests* (1961). The changing paradigms of linguistics from the structuralist-behaviourist period of the 1930- 60s, passing through the psycholinguistic-sociolinguistic period of the 1960- 70s up to the communicative period of today, have been reflected in language tests (Spolsky 1995). The models of language most influential on language testing of the last two decades are the models of communicative competence. This is also the view of language upon which Språkprøven is based. Because of the major influence of communicative competence (CC) on modern language testing, the understanding of models describing the construct is crucial when doing research on modern language tests. In this chapter I shall present the historical evolution of the concept of CC. Different models of CC are presented, concluding in the models of communicative language ability as presented by Bachman 1990 and Bachman and Palmer 1996.

2.1. Communicative competence: Historical evolution of the concept.

The concept of *communicative competence* (CC) is used in language pedagogy, language theory and language testing in a variety of ways and with varying meanings (Savignon 1997:7). CC is defined as "the ability not only to apply the grammatical rules of a language in order to form grammatically correct sentences but also to know when and where to use these sentences and to whom" (Richards et al 1992:65). In the list of language testing terms of the Association of Language Testers of Europe (ALTE) it is defined as "the ability to use language appropriately in a variety of situations". The central importance of appropriateness is also stressed in Crystal's definition of the concept (1997:73). The concept of communicative competence is frequently used in language teaching and language testing. But when does the concept appear in the literature and where does it come from?

2.1.1 Chomsky: Competence versus performance

Chomsky's "rapid and radical success in restructuring linguistics" (Harris 1993:28) is sometimes referred to as "The Chomskyan Revolution". Its impact on linguistics was just as dramatic as any revolution of the social or political kind on society, and it is probably fair to say that every innovation in the field after the 1960s may be seen as either an opposition to or a continuation of Chomsky's ideas. The evolution of the concept of CC is a good example of this.

Chomsky confronted the established paradigm of the structuralist-behaviourists on the level of *what language is* (the linguistic system) as well as on the level of *how it is learned* (language acquisition). The structural-behaviourist linguistics, associated especially with the work of Sapir and Bloomfield in the USA of the 1920s to 50s, viewed language as a finite set of structures and the primary aim of the linguistic enterprise was to describe the structures of particularly the sound and morphological systems apparent in language production (Bloomfield 1914, 1933, Sapir 1949). The approach was based on behaviourist psychology, which claimed that knowledge and beliefs as well as our actions are all the products of rewards and punishment. Language was considered a kind of behaviour, and just as other kinds of behaviour, it was assumed to develop in the child as response to stimuli. Consequently, the language learner was regarded as a passive recipient of linguistic stimuli and language learning as formation of language habits (Skinner 1957).

Chomsky's rejection of the established paradigm rested firmly on what is referred to as *the poverty of the stimulus argument*[2]. Chomsky claimed that the linguistic input that (at least some) children get is not sufficiently rich to account for the fact that all normal children achieve a perfect mastery of the grammar of their mother tongue. There are at least three problems in relation to the input according to the Chomskyan view: It is underdetermined, degenerate and it lacks negative evidence (1989:5). The first of these, the *underdetermination* of input, refers to the fact that the complexity of the grammar acquired by children goes beyond the sentences the individual child may happen to have been exposed to. Children are capable of creating sentences that they have never actually heard in the input. The second problem with the input is its *imperfection*: When we speak we make mistakes, we hesitate, change our minds in mid-course, start over again etc. If the child acquires language only on the basis of the input, one would

[2] Also called *the logical problem of language acquisition, the learnability problem* or *the projection problem* (Chomsky 1981, Larsen-Freeman & Long 1991).

assume them to be confused and sometimes establish language habits which violates the rules of the language they are about to learn. This, however, does not seem to happen. The third problem with the input is its *lack of negative evidence*: L1-research claims to have found empirical evidence that children do not normally get corrected on linguistic form and that when they do, they ignore it (White 1989:14). Then, how do children learn which sentences are not grammatically acceptable? Taken into account the mismatch between the faulty input and the tremendous complexity of the grammatical system achieved by the child, Chomsky argues that language learning would only be possible if the grammatical system were part of children's mental equipment. In other words, human beings as a species must have a genetic predisposition for language learning. In this approach the language learner is not considered a passive recipient of linguistic stimuli, but rather a creative researcher, testing and rejecting hypotheses about the language that surrounds him. The number of possible hypotheses the language learner needs to try out is restricted by the mental grammar as it contains information about ungrammaticality as well about grammaticality.

Chomsky also rejected the structuralist view of what language is. He argued that language is knowledge, not behaviour, and his linguistic theory is one of mental knowledge of a universal grammar (UG) guiding all languages.

The Universal Grammar is defined as" the system of principles, conditions, and rules that are elements or properties of all human languages not merely by accident but by necessity […]" (Chomsky 1976:29). The mental grammar contains *principles*, which are rules applying to all languages and which therefore account for similarities between the languages of the world, and *parameters* that vary within defined limits and which therefore account for language differences. Input is necessary as it contains triggers for the setting of parameter. Once the parameter is set, however, it bears consequences for the language as each setting is assumed to imply a cluster of specific grammatical consequences.

UG with its universal principles and limited set of parameters to be set explains how the child acquires a perfect mastery of the L1 grammar independent of the faulty quality of the input.

> "[…] instead of selecting a rule from a space of infinitely many rules of some rule writing system, the child simply sets the value of an open parameter in some rule already given in Universal Grammar, and thereby derives a language particular rule" (Williams 1987:viii).

The specific proposals of the principles and parameters of the UG are outlined in the Government and Binding (GB) theory (Chomsky 1981, 1986a, 1986b). The theory will not be described in further detail here.

A central dichotomy in Chomsky's linguistic theory is that of linguistic *competence* and linguistic *performance*. According to Chomsky, competence is the perfect knowledge each L1-user possesses of his own language, while performance is the imperfect application of this knowledge in actual language use:

> We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations). […] In fact, [performance] obviously could not directly reflect competence. A record of natural speech will show numerous false starts, deviations from rules, changes in mid-course, and so on (Chomsky 1965:4, author's emphasis).

This distinction between competence as the underlying knowledge of language and performance as the application of this knowledge in language use, is sometimes referred to as a the weaker claim of Chomsky's concept of competence (Canale and Swain 1979:4).

According to Chomsky the focus of linguistic theory should be on competence, not on performance:

> Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance (Chomsky 1965:3).

Chomsky's stronger claim is the assertion that the concept of linguistic competence should be restricted to the tacit knowledge of the grammatical system alone. According to this view a theory of competence is equivalent to a theory of grammar, and the linguistic theory launched by Chomsky is indeed a theory of this kind.

Most linguists seem to accept the weak claim of the competence/ performance dichotomy (Canale and Swain 1983:4). The strong claim, however, is met with restive opposition in the 70s, especially from the camp of the sociolinguists.

2.1.2 Hymes: Communicative competence

The concept of *communicative competence* was introduced by Hymes in 1972 as a reaction to and, at the same time, extension of the Chomskyan concept of competence. Hymes criticised the Chomskyan concept for being to narrow in scope. Obviously, there are more to language than grammar, hence there must be more to linguistic competence than the knowledge of grammar.

> We have then to account for the fact that a normal child acquires knowledge of sentences, not only as grammatical, but also as appropriate. He or she acquires competence as to when to speak, when not, and as to what to talk about with whom, when, where, in what manner (Hymes 1972:277-78).

In other words, "There are rules of use without which the rules of grammar would be useless" (Hymes 1979:15).

The Hymesian approach includes "several sectors of communicative competence, of which the grammatical is one" (Hymes 1979:18). The four sectors refer to whether or not something is possible, feasible, appropriate, and in fact done. Something is *possible* if it is acceptable on a grammatical, cultural or communicative level. *Feasibility* has to do with restrictions not necessarily linguistic, such as "memory limitations, perceptual devices, effects of properties such as nesting, embedding, branching, and the like" (Hymes 1972:22). Whether or not something is *appropriate*, is a key term in the Hymesian concept of CC. Indeed, Hymes claims, the judgement of which utterances are appropriate and which are not, requires a tacit knowledge on the level of competence, even in a Chomskyan sense of the word. A sentence may be grammatically correct but inappropriate in a given context. Successful use of the language requires knowledge of appropriateness as well as of grammaticality. The final sector is whether something is actually *done*. An utterance may be possible, feasible and appropriate but still not performed.

Summing up then, Hymes is credited for introducing the term of communicative competence, and for expanding the Chomskyan concept to include aspects of appropriateness and the ability to use language competence in actual communication.

2.1.3 Canale and Swain

Canale and Swain (1980) develop a model of communicative competence building on the wider use of the term introduced by Hymes. As Hymes they argue that grammatical competence is part of the knowledge about language that a language users possesses, and that there is more to language competence than the knowledge of grammar. As distinct from Hymes and Chomsky,

their primary focus is on the second language learner. Canale and Swain offer an (integrative) theory of CC:

> […] in which emphasis is on preparing second language learners to exploit – initially through aspects of sociolinguistic competence and strategic competence acquired through experience in communicative use of the first or dominant language – those grammatical features of the second language that are selected on the basis of, among other criteria, their grammatical and cognitive complexity, transparency with respect to communicative function, probability of use by native speakers, generalizability, to different communicative functions and contexts, and relevance to the learners' communicative needs in the second language" (Canale and Swain 1980:29).

Their model of communicative competence includes three main competencies: grammatical, sociolinguistic and strategic competence. The *grammatical competence* is taken to include knowledge of lexical items and rules of morphology, syntax, sentence-grammar semantics, and phonology. The *sociolinguistic competence* is assumed to include two sets of rules: sociocultural rules and rules of discourse. The sociocultural rules govern the production and interpretation of utterances in relation to what is appropriate within a sociocultural context depending on factors such as topic, role of participants, setting, and norms of interaction. The sociocultural rules also govern the choice of the appropriate attitude and register. The other kinds of rules included in the sociolinguistic competence are *the rules of discourse*: Canale and Swain acknowledge that this category lacks a precise definition, but they assume it to include the rules of cohesion and coherence. It is not clear how these rules are distinct from the grammatical rules governing cohesion and the sociolinguistic rules governing coherence, though. (Indeed, the problematic status of this category is one of the causes for Canale's revision of the model in his 1983 article presented below). The final component of the model of CC, is the *strategic competence*. Canale and Swain define this as the "verbal and non-verbal communication strategies that may be called into action to compensate for breakdowns in communication due to performance variables or to insufficient competence" (1979:56).

2.1.4 Canale 1983

In his 1983 article Canale adds to the model of CC one extra component, that of *discourse competence*. Or rather, the rules of discourse which in the 1980 model were classified as a subcategory of sociolinguistic competence, are given the status of one of the main competencies in the revisited model. Hence, according to Canale 1983, CC includes four main competencies: grammatical competence, sociolinguistic competence, discourse competence and strategic competence. *Discursive competence* is defined more precisely than in the 1980 model as the "mastery of how to combine grammatical forms and meanings to achieve a unified spoken

or written text in different genres" (1983:9). Unity of a text is achieved through cohesion in form and coherence in meaning. The *strategic competence* is also given a broader meaning in Canale's model: It is taken to compensate for breakdowns in communication as in the earlier version of the model, but in addition it is assumed to enhance the effectiveness of communication.

2.2 Bachman 1990: Communicative language ability

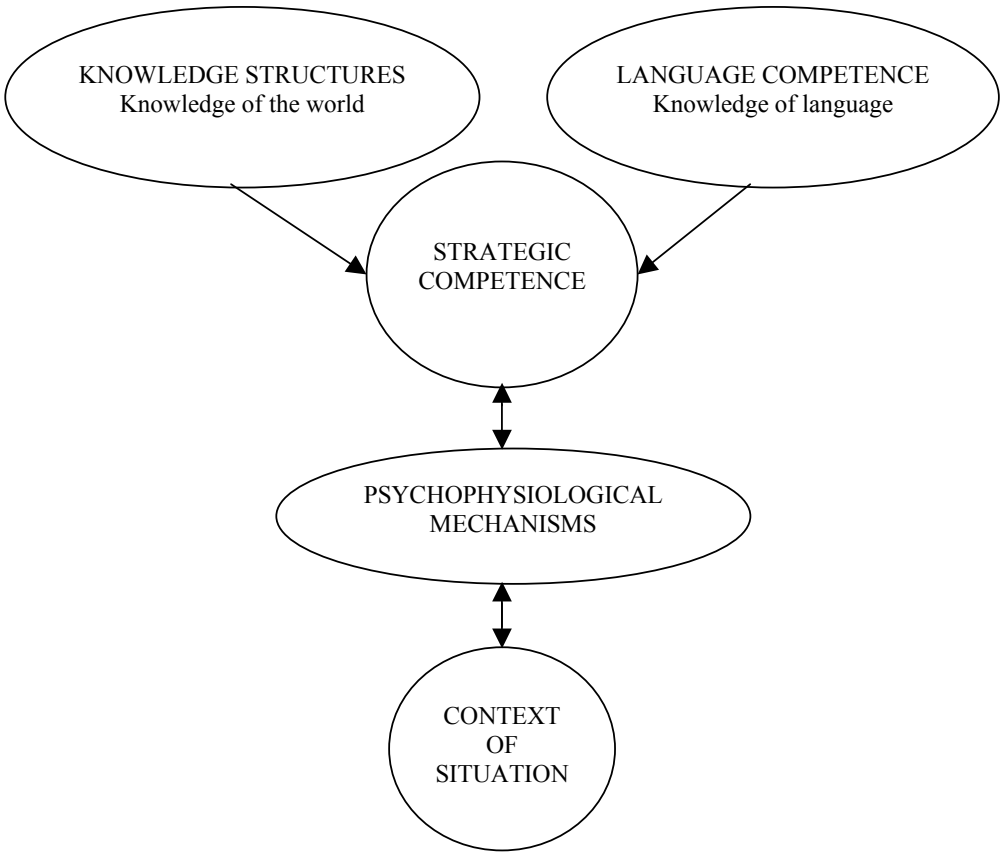
The theoretical model of communicative competence presented by Bachman in 1990 and further developed by Bachman and Palmer in 1996 introduces a measurement perspective to the earlier models of CC. Indeed this model is the one underlying most of today's professional language testing (Saleva 1996). The test in focus of the present study, "Språkprøven i norsk for voksne innvandrere" is also grounded on this theoretical framework. The model therefore deserves a treatment in some detail.

Though based on linguistic research, Bachman's 1990 model has evolved through empirical research in language testing (Bachman 1990:82). Through the use of multitrait-multimethod (MTMM) design (Campbell and Fiske 1959) and confirmatory factor analysis, Bachman and Palmer investigated the nature of language proficiency on the basis of test performance. Their research offered two important findings: It led to a revision of the model of CC as presented by Canale and Swain 1980 and Canale 1983. In addition, their research affirmed the assumption that the results on a language test are affected not only by test takers' language proficiency, but also by individual characteristics of test takers and test method facets (Bachman 1990:37, Harley et al 1990:37). This is an important acknowledgement underlying language testing, and it will be further treated in Chapter 4. Here we shall focus on the model of CC, which grew out of this research.

The theoretical framework proposed by Bachman 1990 comprises both the knowledge of language and the capacity for implementing that knowledge in actual communication. It includes three components: *language competence, strategic competence* and *psycho-physiological mechanisms*. The language component entails components of language competence similar to the models of CC as described by Canale and Swain 1980 and Canale 1983. Strategic competence is by Bachman characterised as "the mental capacity for implementing the components of language competence in contextualized communicative language use" (Bachman 1990:84). Psycho-physiological mechanisms refer to the psychological and physical processes involved in the

actual execution of language. These components interact with other components of the language user's knowledge (knowledge of the world) and with the context of situation in which language occurs. The components of CLA are represented in Figure 1 below:

Figure 1 Components of communicative language ability (CLA) in communicative language use (Bachman 1990:85).



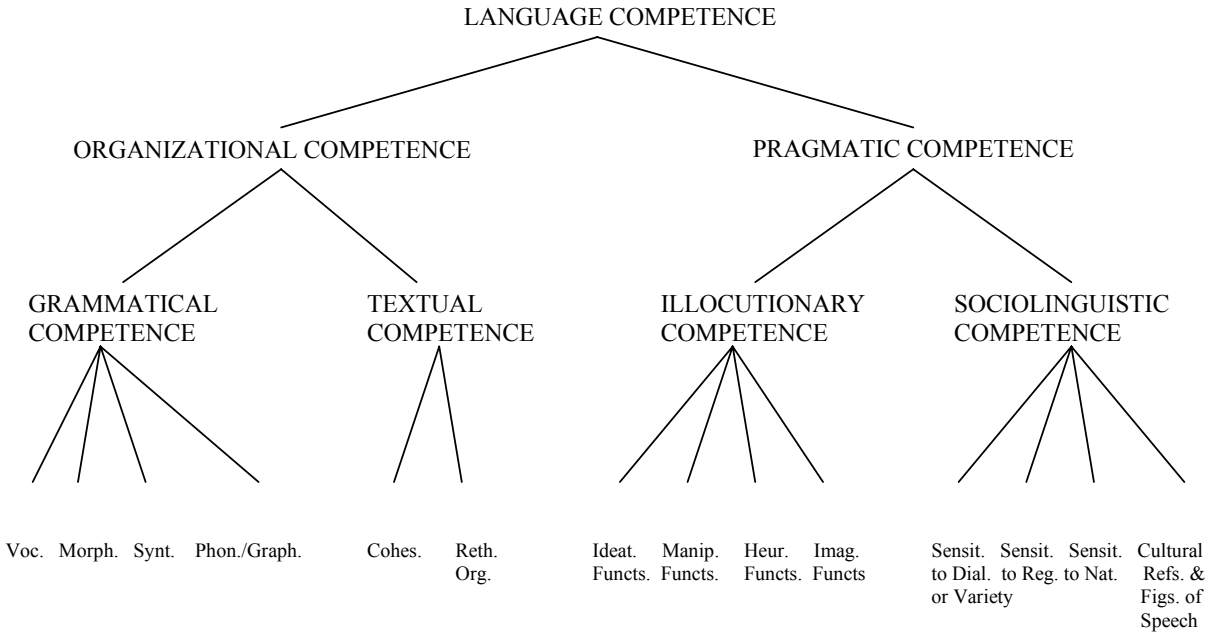
Bachman's term *language competence* covers to some extent what in earlier models was named communicative competence. One important difference between Bachman and Canale's models is their treatment of *strategic competence*. Bachman sees strategic competence not as a language-specific device, but as a general ability applying to non-language situations as well:

> "I consider [strategic competence] more as a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task, whether that task be related to communicative language use or to non-verbal tasks such as creating a musical composition, painting, or solving mathematical equations" (Bachman 1990:106).

Unlike Canale, Bachman sees strategic competence as crucial in all language use, and not only as a means to solve a communicative problem or to compensate for an insufficient language control. He includes three components in strategic competence: an assessment component, a planning component and an execution component. The *assessment component* enables the language user to identify the information needed, the language competencies available, the knowledge shared by the interlocutors and finally the extent to which the communicative goal has been achieved. The *planning component* enables the language user to retrieve the relevant items from language competence according to a plan of how to reach the communicative goal. Finally, the *execution component* relates to the psycho-physiological mechanisms to implement the plan in the appropriate channel (oral or visual) and mode (spoken or written).

The remaining components of language competence of Bachman's model are visualised in Figure 2:

Figure 2 Components of language competence (Bachman 1990:87).



Language competence comprises two main categories, *organisational* and *pragmatic competencies,* which may be further divided in subcategories. The *organisational competence* consists of those

abilities involved in controlling the formal structure of language both on sentence- and text level. These abilities are of two kinds: grammatical and textual. The *grammatical competence* includes relatively independent competencies such as the knowledge of vocabulary, morphology, syntax and phonology and graphology. The *textual competence*, on the other hand, comprises the knowledge necessary for joining utterances together to form a written or oral text. It is of two kinds, cohesion and rhetorical organisation. *Cohesion* refers to the way semantic relationships are established grammatically through the use of reference, ellipsis, conjunction, and lexical cohesion as well as those governing the presentation of new and old information. *Rhetorical organisation* refers to the overall structuring of the text and is related to the intended effect of the text on the reader/ listener. It includes methods of development such as narration, description, comparison, classification and process analysis. In oral conversation it includes the competencies involved in the organisation and performance of turns such as attention getting, topic nomination, topic development and conversation maintenance (Hatch 1978) as described in discourse analysis.

The other main component of language competence is *pragmatic competence*. This ability relates to "the relationship between utterances and the acts or functions that speakers (or writers) intend to perform through these utterances […] and the characteristics of the context of language use that determine the appropriateness of utterances" (Bachman 1990:90). Pragmatics, then, refers not to whether utterances are grammatically correct and coherent, but whether they are appropriate, acceptable and successful in relation to the intended meaning of the language user.

Pragmatic competence comprises two main components: illocutionary competence, on the one hand, and sociolinguistic competence on the other. Bachman describes *illocutionary competence* by reference to the theory of speech acts on the one hand and by reference to language functions on the other. Searle 1969 distinguishes between three kinds of speech acts: utterance acts, propositional acts and illocutionary acts. An *utterance act* is simply the act of saying something. A *propositional act* involves referring to or expressing a predication about something. An *illocutionary act* is the function performed in saying something, such as asserting, warning, requesting etc. The illocutionary force is the communicative intention of an utterance. Central in theories of illocutionary competence, is the understanding that the illocutionary force of an utterance is independent of its grammatical form or sentence type. Take the following dialogue between a young man and woman in the cold Norwegian winter night:

A: Are you cold?
B: Yes, please!

A 75 year-old man told me this as an answer to how he met his wife to whom he had been married for 50 years. Obviously, she had no problem in understanding the illocutionary force of his statement being a request. She understood that what he *really* meant but was too shy to spell out directly, was if she would like him to hold her tight. The example illustrates that the illocutionary competence is used in producing as well as in interpreting utterances. A more commonly referred example is the statement "It is cold in here" which may be given different meanings such as: "Turn on the heat!/ Close the window!/ Don't bring the baby in here! or again: "Please hold me tight!" depending on the context.

Bachman also describes the illocutionary competence by reference to the description of language functions as presented by Halliday (1973, 1976). Bachman divides language functions in four main groups: ideational, manipulative, heuristic and imaginative. The *ideational* functions are the most common use of language. It means the simple expression of meaning and exchange of information about knowledge of the world, feelings, thoughts etc. (Halliday 1973:20). The second group of functions comprises the *manipulative* functions in which the purpose is to affect the world around us. The manipulative functions may be of different kinds, I will limit this overview to mentioning the *interactional* function which is the function of language "to form, maintain, or change interpersonal relationships" (Bachman 1990:93). The main function of much interpersonal language is the maintenance of a social relationship rather than the conveyance of information. This is the obvious case for phatic language use, such as greetings, ritual inquiries about health, or comments about the weather (discussed in more depths in the next chapter). The *heuristic* functions of language are those in which the purpose is to extend our knowledge of the world around us, or even our knowledge of language itself. It is naturally a common function in teaching and learning, but also as part of everyday conversations. The *imaginative* function of language "enables us to create or extend our own environment for humorous or esthetic purposes, where the value derives from the way in which the language itself is used" (Bachman 1990:94), exemplified by telling jokes, constructing and communicating fantasies, creating new metaphors as well as using language creatively in reading or writing literary works.

The other main component of pragmatic competence is the *sociolinguistic competence*. A central aspect in relation to this component is whether or not something is appropriate, a cardinal aspect in all post-Chomskyan definitions of CC as already mentioned. Bachman defines sociolinguistic competence as:

> […] the sensitivity to, or control of the conventions of language use that are determined by the features of the specific language use context; it enables us to perform language functions in ways that are appropriate to that context (Bachman 1990:94)

The sociolinguistic sensitivity, or control, may be of distinct kinds: It entails the sensitivity to different dialects or varieties, to understand and classify differences in register and naturalness, as well as the ability to interpret cultural references and figures of speech.

Sensitivity to differences in dialect or variety refers to a language user's ability to judge which dialect or language variety would be appropriate in a given context. This, of course, varies from one language society to another. In Norway there is a relatively high tolerance for dialectal variation as compared to other language communities and people speak their local dialects on television and radio as well as in a lecture at the university. The case is different in for example England and France where it would probably be more appropriate to use a standard variant of the language in such contexts.

Sociolinguistic competence also involves *sensitivity to differences in register*. Bachman builds on Halliday, McIntosh, and Strevens (1964) who define register as the kind of variation in language use within a single dialect or variety, which relates to the field, mode or style of discourse. The *field of discourse (*or discourse domain) may refer to the subject matter of the discourse (in lectures, discussions or written expositions etc.) or to the language use context (commenting a football game, computer-language etc). Different subjects and different contexts demand a different language register. The *mode of discourse* (written or spoken) also affects the register used. When a written text is presented orally we react to it as stiff and formal, when an authentic dialogue is transcribed, it strikes us as disjointed and chaotic. (These differences between the written and spoken mode will be treated in more detail in the next chapter). The third aspect of differences in register, is the *style of discourse,* which relates to the relations among the participants. Joos (1967) distinguishes five levels of style: frozen, formal, consultative, casual and intimate. As these styles reflect the degree of intimacy between the participants of the discourse, the choice of an inappropriate style may be interpreted as rude if a too familiar style is chosen, or odd and even comic if a too formal style is used.

Another aspect of sociolinguistic competence is the language user's *sensitivity to naturalness*. An utterance may be grammatically correct and coherent, it may fulfil its illocutionary purpose as well as being appropriate when it comes to dialect or register, and still seem unnatural in the language context. This is often the case for advanced second language learners who have gained control of the formal as well as the pragmatic aspects of language. Yet some of their utterances, correct as they may be, just sound unnatural. My French speaking fellow students at a French university had great fun on my behalf because of the unnaturalness or non-nativelikeness of the

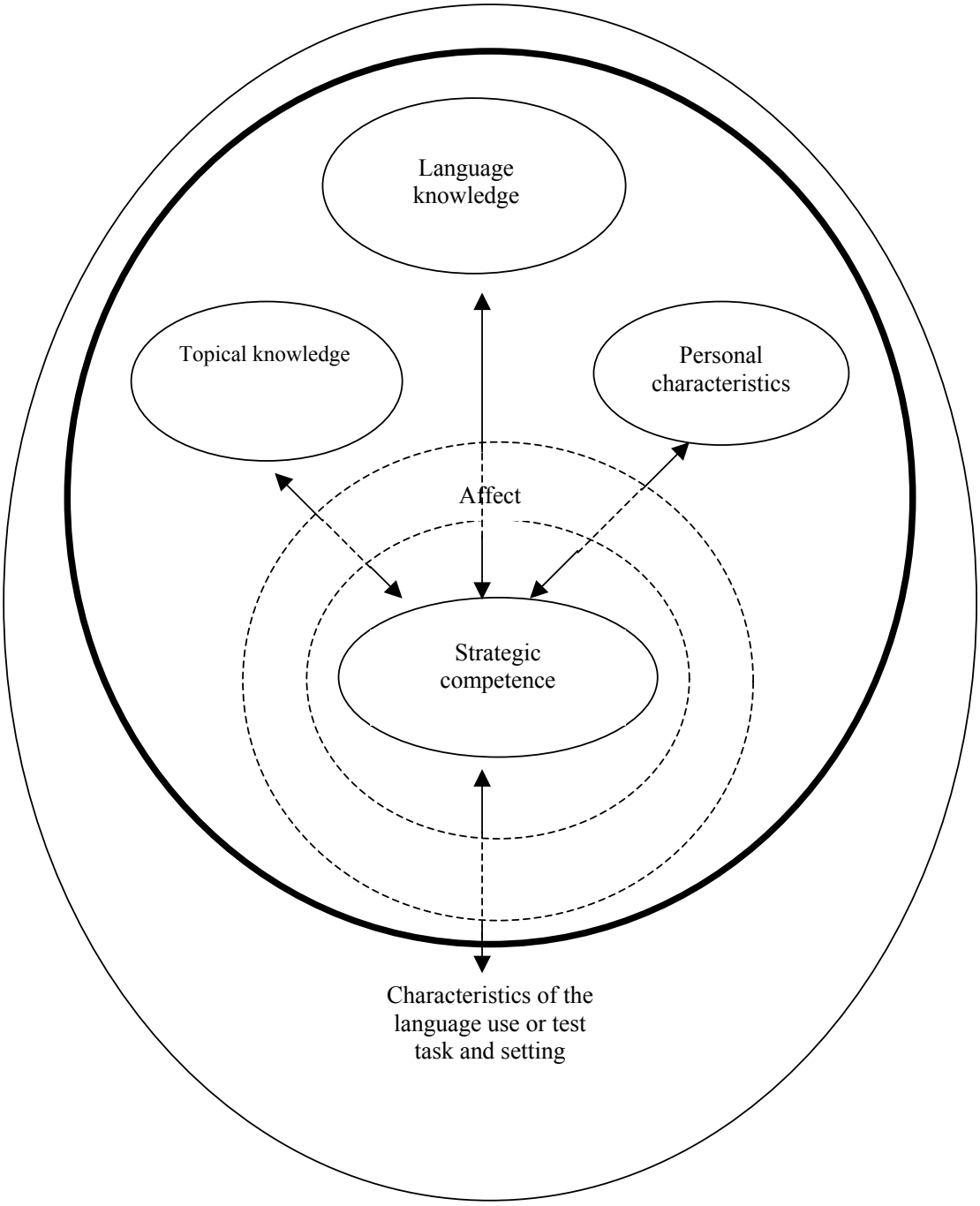
French I had acquired at a Norwegian university. Apparently the study of Molière's plays had been on the expense of practice in modern French conversation: I used words and expressions that had not been used in the streets of France for centuries.

The final aspect of sociolinguistic competence mentioned in Bachman's model is the *ability to use and interpret cultural references and figures of speech*. A widely used *cultural reference* in Norwegian society is: "Det blir som å hoppe etter Wirkola"/ "It will be like jumping after Wirkola". Understanding the expression requires knowledge of the famous Norwegian ski jumper of the 1960s, Bjørn Wirkola. Today the expression is used not with reference to ski jumping at all, but whenever one's performance is preceded by an outstanding performance. Understanding *figures of speech* also requires more than an understanding of the meaning of the words per se, and as cultural references they are often different from one language to another. The meaning of the Norwegian expression "å synge på siste verset"/ "to be singing on the last verse" is not obvious for a non-native speakers even if he does understand each separate word. The meaning would be expressed by the use of a different figure of speech in English: "to be on its last legs". There are hundreds of examples of figurative language use. Some figures are specific for only one language society while other are shared by a larger society, such as the western world, the Arab world etc. Because of their tight connection to the cultural and social surroundings of language, Bachman places the ability to use these devices under sociolinguistic rather than organisational and vocabulary competencies.

2.3 Bachman and Palmer 1996

In Bachman and Palmer 1996 the model presented by Bachman 1990 is revisited and some changes are made. In the revisited model there is a greater emphasis on individual characteristics of the test-taker and it is more closely related to the language test situation than the earlier model:

Figure 3 Components of language use and language test performance, Bachman and Palmer 1996: 63.



Strategic competence keeps its central position from the 1990 model, and as evident in the figure above it is taken to interact not only with language but with topical knowledge and personal characteristics as well as with the situation in which language is used. Topical knowledge replaces the term knowledge of the world of the earlier model. Personal characteristics are added and language use is related to the language use situation especially the test situation. (The component of psycho-physiological mechanisms is not explicit in the revisited model).

Bachman and Palmer list four kinds of *individual characteristics,* which should be taken into consideration when constructing language tests.

1 personal characteristics, such as age, sex, and native language,
2 the topical knowledge that test takers bring to the language testing situation,
3 their affective schemata, and
4 their language ability

(Bachman and Palmer 1996:64)

Personal characteristics are individual attributes other than the candidates' language ability but which nevertheless affect their performance on a language test. Such characteristics are age, sex, nationality, resident status, native language, level and type of general education, type and amount of prior experience with a given test as well as socio-psychological factors, personality, cognitive style, foreign language aptitude etc.

The *topical knowledge* refers to the individuals' factual knowledge about a given topic. As underlined by Bachman and Palmer, a test taker's performance on the tests does not only depend on his or her language ability. Their knowledge of the topic in question may affect their performance negatively or positively.

The *affective schemata* refer to the way that test takers respond emotionally to the test tasks. As the topical knowledge mentioned above, the affective schemata might facilitate or inhibit performance: If test-takers feel comfortable, at ease or inspired by the test tasks they may perform better. If they are nervous, provoked or intimidated by topics raised in the tasks the opposite may be the case.

When it comes to the treatment of *language ability* the differences between the 1990 and the 1996 models are mainly on a terminological level. The term communicative language ability or CLA in the 90-model is replaced by the term *language use* in the revisited model. The language component, which in the earlier model was called language competence, is replaced by the term *language knowledge*. This replacement of the term competence with the term knowledge is implemented throughout the entire model (with the exception of strategic competence), so that

in the 96 model we find *organisational* and *pragmatic knowledge* with their subcategories *grammatical* and *textual knowledge* on the one hand and *functional* and *sociolinguistic knowledge* on the other. Functional knowledge replaces the earlier term of illocutionary competence.

In Bachman and Palmer 1996 the treatment of *strategic competence* is more tightly connected to the test situation than in the earlier version of the model. It is still divided into three components (or areas, in the 1996 terminology) but these are somewhat different from the earlier model: The three areas are goal setting, assessment and planning (replacing the assessment, planning and execution components of Bachman 1990). The *goal setting area* involves deciding what to do, identifying the test tasks and choosing between them. The *assessment area* includes deciding what is needed to fulfil a task, what language recourses and topical knowledge is available, as well as evaluating how well one has done. And finally, the *planning area* involves the decision of how to use the recourses available. It means selecting the relevant items of one's language ability and topical knowledge and selecting a plan for implementing these elements in a response to the test task. When comparing the treatment of strategic competence in the 1990 and 1996 models, then, the differences can be summarised as follow: The term goal setting area is new in the later model. So is the content of the term. The assessment area conforms to the content of the assessment component in the earlier model. The term planning area covers the planning as well as the execution components of the earlier model. Apart from some terminological differences between the treatment of strategic competence in the two models, then, the main difference is that in the later model, it is more closely connected to the test situation. In the LT literature frequent references are made to both Bachman 1990 and Bachman and Palmer 1996. The latter does not seem to have completely replaced the former. In this thesis, I build mainly on the 1996 model.

Summing up the treatment of communicative competence, then, Chomsky (1965) drew the line between linguistic competence and linguistic performance. The Chomskyan concept of competence was restricted to the knowledge of the grammatical system. Hymes (1972) expanded the concept of competence to include sociolinguistic competence as well. Canale and Swain (1980) and Canale (1983) described a model of CC in relation to second language learning consisting of various components, a model which was used as a basis for empirical research by Bachman and Palmer in their investigation into the nature of CC (1982). Bachman 1990 and Bachman and Palmer 1996 related CC to language testing in developing the models of CC. These later models of communicative language are generally accepted as a theoretical base,

or construct, of performance testing in modern LT. A problem with these models, however, is that they do not distinguish between written and oral language use, a point also made by Saleva (1997).

CHAPTER 3: SPEAKING

To ensure that a test measures what we think it measures, it is of major importance that we know the skill we are testing and that we are able to describe it in detail (Messick 1975, Bachman 1990). In the previous chapter I argued that the models of communicative competence, influential as they have been on modern language test construction, fail to distinguish between the spoken and the written mode. Bachman of course acknowledges the differences between the two modes of expression, but nevertheless argues that the conventions guiding conversational language use can be best described in a general model of CC (Bachman 1990:89). However, these models have been criticised for being so general that they are almost impossible to operationalise for test purposes (Saleva 1997:18). It is my opinion that the general models of CC are useful but insufficient for a thorough understanding of what spoken language is. In a project like the present focusing on the measurement of speech, more specific models are needed, and hence, the main purpose of this chapter is to describe presumed characteristics of spoken language and thereby introduce a terminological framework that will be used later in the thesis.

The chapter begins with an overview of the components of speaking building mainly on Bygate (1987). Thereafter formal differences between speaking and writing are dealt with (in Section 3.1). These differences are seen as a result of two conditions affecting the spoken language: the fact that speaking occurs under the pressure of time on the one hand, and that it (normally) occurs between interlocutors, on the other. In section 3.2 functional differences between speech and writing are handled: while the main function of writing is the transmission of information, it is argued that the principal function of the spoken language is to establish and maintain social relationships.

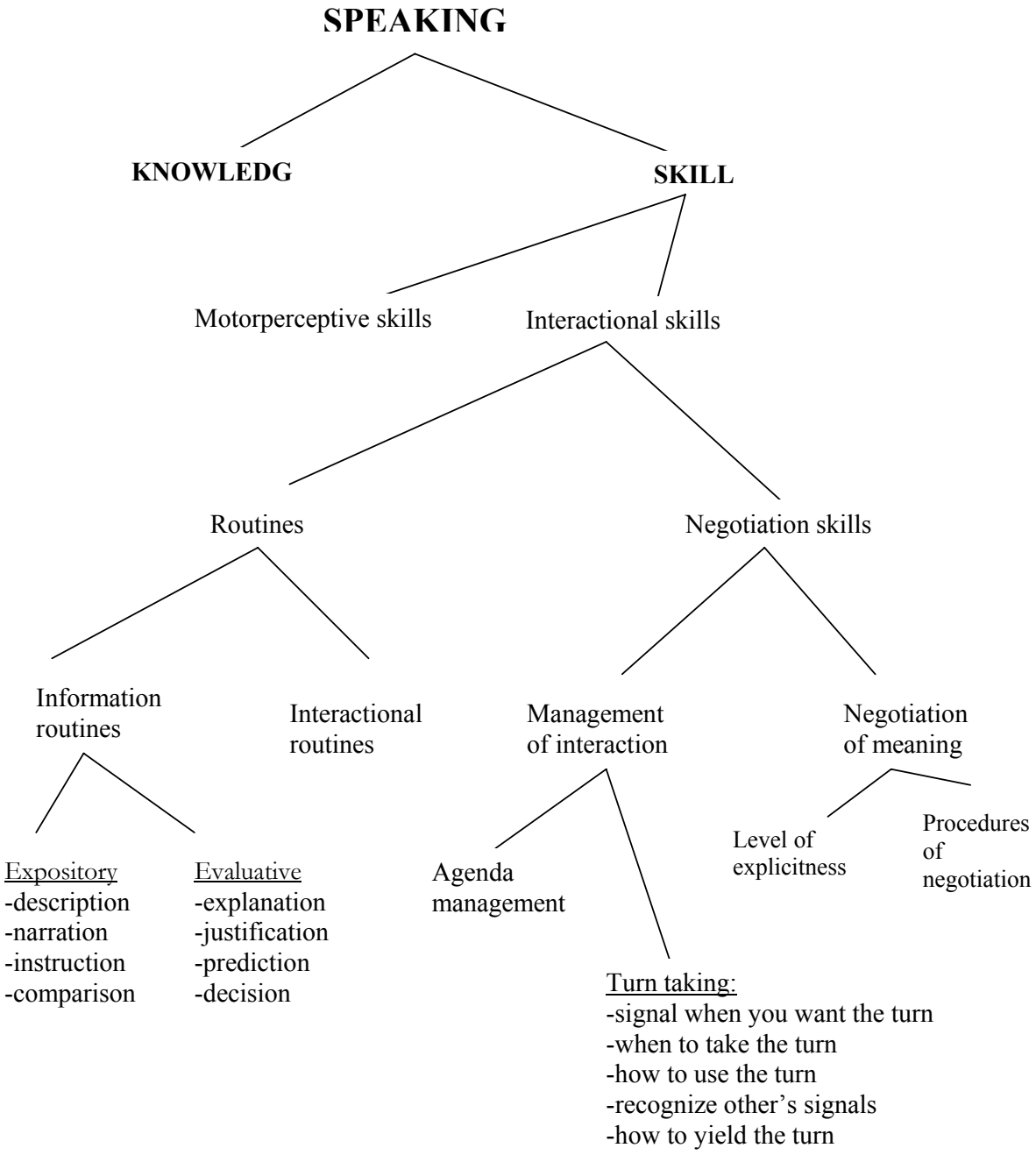
In this chapter the main focus is on the spoken language of native language users. Language testing is a normative science: the language proficiency of test candidates is traditionally compared to some norm of correctness. LT therefore takes greater interest in describing this norm than in describing the language of learners which, together with describing language acquisition, belongs to the field of second language acquisition (SLA) research. It is however worthy of mentioning that in modern LT this tradition is being challenged. Several scholars suggest that rating scales of L2-tests describing different proficiency levels should be based on, or validated against, the performance of language learners (Saleva 1997, Fulcher 1996a, Brindley 1998). The attempts of using models of language acquisition such as Processability Theory in this respect are promising. This will be discussed in some detail in relation to rating scale

development in Chapter 6. This thesis is however written in relation to one specific test of Norwegian as a second language, which is traditional in that its rating scale descriptors are normative: the candidates' language is not described in its own right in relation to stages of acquisition. Even though I find the idea of basing rating scale descriptors on knowledge of language acquisition very tempting, I will not follow it further here, but restrict the overview of speaking to a general description of this skill in comparison with the skill of writing. When looking at oral rating scales, one may sometimes get the impression that speech is compared to a norm of written language more than with actual speech production. When oral rating scales refers to grammatical correctness, precise and varied vocabulary etc. one may suspect that test constructors have a written norm in mind. One of the main purposes of this chapter is therefore to describe characteristics of the oral mode, which is a prerequisite in order to develop and validate rating scales for L2 speech.

3.1 An overview of the components of speech

In his book *Speaking* (1987) Bygate offers a detailed description of the composite skill of speaking. The tree-diagram in Figure 4 below is a graphical representation of Bygate's verbal overview.

Figure 4 Bygate's overview of speaking, graphical representation.



Bygate makes a fundamental distinction between *knowledge* of the different elements of speech (grammar, pronunciation, vocabulary, pragmatic knowledge etc.) and the ability to implement this knowledge in actual language use: i.e. the *skill* of speaking. Bygate clarifies the difference between knowledge and skill, by using a practical example of driving a car. There is a great difference between knowing which pedal is the brake and which is the accelerator, how to put the car in reverse, etc. and the actual skill of driving the car. It is helpful to know the function of the different controls of the car, but it is not obvious that a good formal knowledge makes you a better driver. So to learn how to drive you need practice in performing the skill of driving. The case is similar for speaking: it is generally accepted that one may have considerable knowledge about a language, one may have studied its grammar in detail, without being able to actually speak it (Krashen 1977, 1978). There is a consensus in the field that language teaching and testing should be concerned with the practical language skills more than with the formal knowledge of its parts.

At another level, Bygate distinguishes between motor-perceptive skills and interactional skills. The *motor-perceptive skills* "involve perceiving, recalling, and articulating in the correct order sounds and structures of the language" (Bygate 1987:5). We are no longer talking about a passive knowledge of the elements of speech, but of a practical use of the different elements in actual speech production. The motor-perceptive skills differ from the interactional skills, in that it is a context-free use of language: it is the skill of producing speech out of context, which was traditional in the audio-lingual approach that dominated language teaching and– testing from the 1930s to the 1960s. Bygate compares it with driving a car on a deserted road far from the flow of traffic.

The most common purpose of the spoken language is interaction or conversation. The *interactional speaking skill* is the implementation of speech in face- to- face communication. It involves making decisions about communication: what to say, as well as when and how to say it. It involves judgement of the interlocutor's prior knowledge about the topic. It comprises the ability to recognise signals from the interlocutor that he is bored, that he doesn't understand, or that he wants to talk himself, for example, as well as the ability to formulate one's message sufficiently precisely for the message to get across without too much detail, which may bore, or even insult the interlocutor. Interactional skills of speech may be compared with driving a car in a

city street with heavy traffic, where you have to manage the controls of the car and at the same time pay attention to signs, pedestrians and cars around you.

In Bachman's model of CC, interactional skills are categorised as part of the general textual competence. Bachman states:

> "These conventions, such as attention getting, topic nomination, topic development and conversation maintenance (Hatch 1978) appear to be ways in which interlocutors organize and perform the turns in conversational discourse, and may be analogous to the rhetorical patterns that have been observed in written discourse" (Bachman 1990:88).

As is evident in this quote, Bachman focuses on similarities over differences between the two modes of expression in his model.

Since the 1970s, when the communicative approach became dominant, the field has been striving to teach and test interactional skills in context. But even though conversation is a common use of speech, it is not unproblematic to measure (Underhill 1987, Bachman 1990, Weir 1990). We shall return to this point in Chapter 6.

The interactional skills may on their part be divided into two groups of subskills: routines (Widdowson 1983) on the one side and negotiation skills on the other. *Routines* may be defined as "conventional ways of presenting information" (Bygate 1987:23) They are similar to what is generally called *genres of speaking* (Section 3.3 below). A genre or routine refers to special kinds of speech which follow more or less fixed schemes of expression such as, for instance, story or joke telling, instructions, interviews, lectures, telephone conversations etc. Once learned and automated the routines may be easily processed and applied in speech without much planning time. Bygate subcategorises the routines into two main groups: information routines and interaction routines. *Information routines* include "frequently recurring information structures" such as stories, descriptions, presentations of facts, comparisons, instructions etc. These may again be divided into two main groups according to the kind of information they convey. The routines that mainly consist of transmitting factual information are called *expository routines*. Brown & Yule (1983) identify three kinds of expository routines: *description, instruction* and *narration.* Bygate adds *comparison* to this list. As opposed to the information routines, the *evaluative routines* involve the drawing of conclusions, usually involving the expression of reasoning. Typical examples of evaluative routines are explanation, justification, prediction and decision. They usually build on the factual information of the expository routines, but this is not an absolute condition.

The other kind of principal routines are *interactional routines* which are "based not so much on informational content as on sequences of kinds of terms occurring in typical kinds of

interactions" (Bygate 1987:25), such as telephone conversations, lessons, television interviews etc. Such routines tend to be organised in characteristic ways. This becomes obvious when normal patterns are violated.

As opposed to routines, *negotiation skills* are common to all kinds of communication. We apply negotiation skills to make ourselves understood as well as to cope with different kinds of communicative problems that may arise during the conversation. Negotiation skills may therefore be compared to communications strategies in Bachman's sense of the word. The negotiation skills may be divided into two subcategories: management of interaction on the one hand, and negotiation of meaning on the other. *Management of interaction* refers to the devices used in a free face-to-face conversation between two or more participants. The participants have to agree about the topics of the conversation, as well as on who is to talk and for how long. The choice, development and shifts of topic are grouped together in what Bygate calls *agenda management*. So is the basic freedom to start, maintain and end a conversation.

In addition to control of the topic, the participants have to know how to handle *turn taking*. According to Bygate, efficient turn taking requires five abilities: first the participants have to know how to signal in an appropriate and polite manner that they want to speak. Second, they need to know when the right moment for taking the turn is, which means recognising other participants' signals. And once a speaker has got the turn, he has to know how to keep and use it appropriately so that he does not lose it before he has managed to get his message across. A fourth ability is to recognise the interlocutor's signal that he wants the turn. Finally, the speaker has to know how to yield the floor to someone else. Failure to recognise such signals may cause frustration and irritation on the part of the interlocutors. It is important to notice that native speakers are not equally good at turn taking. Some find it hard to get the turn when they want to, some find it hard to express what they want in efficient ways, and some have problems in recognising other people's signals and yielding the floor in time (Tannen 1984). Odlin (1989) claims that there is considerable cross-linguistic variation within the rules governing turn-taking. The language-specific rules of turn- taking may cause problems when learning a new language.

The other component of negotiation skills, is *negotiation of meaning*. This entails the ability to get one's message across successfully. In contrast to writing, getting the message across depends not only on the speaker but on the listener as well. The speaker has to take the interlocutor's background knowledge and linguistic skills into account to choose an appropriate *level of*

explicitness. Too much or too little information may both disturb the flow of information. Even native speakers do sometimes misinterpret the background knowledge of their interlocutors and fail to give information at the right level of explicitness.

In addition to choosing an appropriate level of explicitness, the speaker uses *procedures of negotiation* to ensure that the interlocutor actually understands the message conveyed. To ensure that the information is at the right level, the speaker uses questions of clarification. In addition the listener may indicate whether he has understood the message by asking for repetition or rephrasing. There is mutual effort between the interlocutors in finding a suitable level of explicitness. If a listener does not help out during the conversation, he may give the impression of being rude and uncooperative (Grice 1989).

So far, we have been dealing with the sub-skills of speaking based on Bygate's categorisation. The different components could probably be categorised differently. Nevertheless, Bygate's overview serves its purpose here, which is to show that speaking consists of much more than knowledge of a language's purely linguistic parts such as vocabulary, syntax, morphology and pronunciation. In the following section we shall look at some characteristics of the spoken, as opposed to the written, language. Firstly, formal differences between speech and writing are discussed. Thereafter we will look at the different communicative functions of the two modes of expression.

3.2 Formal differences between writing and speaking.

Written language has generally been more appreciated than spoken language, and we still know more about the rules governing written language than about the rules governing speech (Bygate 1987:vii, Stenström 1994:xi). For the written language, we have models and formal rules spelled out, while the situation for speaking is different. Not all languages have a pronunciation norm like the English Received Pronunciation (RP). The Norwegian speech community, for instance, has a variety of different dialects, which generally hold a high status. It is not evident which dialect should stand as a norm in, for instance, foreign language teaching. Rather teachers are recommended to teach the spoken language of the local environment of the students. The spoken language varies a lot more than the written language. In addition to the dialectical differences, spoken language varies according to sociolinguistic differences. A young man uses a quite different sociolect than an elderly lady within the same dialect area. A great deal of variation in combination with the lack of a formal norm and description of rules governing speech complicate the teaching and testing of spoken language. In addition, the myths that the spoken

language may be taught simply through providing spoken input, has not helped in the urge for more knowledge about this skill, a point also made by McLaughlin (1987).

In the literature, speaking is often treated simply as written text spoken aloud. But even though there are similarities between speaking and writing, a written and spoken text do indeed have their distinctive features. This fact is immediately appreciated when someone presents a written text orally without giving it a more oral shape. We react to it as stiff and formal, too complex for processing and tiring to listen to. Similarly, the transcription of an oral conversation strikes us as disconnected, imperfect and chaotic. Still it makes perfect sense to us when we speak. So clearly there must be greater differences between writing and speaking than only the mode of expression. In the following sections the characteristics of writing and speaking will be discussed and some possible explanations of their characteristics will be investigated.

3.2.1 Conditions affecting oral language use

Writing and speaking normally take place under quite different circumstances. When we write, we usually have time to plan both the content and the form of our message. We have time to choose a correct and appropriate language, and even to make sure there is enough variation in syntax and vocabulary to prevent the text from being dull and monotonous to read. If we make mistakes, we may correct them before we deliver the text. For the spoken language, the situation is quite different: when speaking, we have to plan what we are going to say as we speak. Because of the time limitation we have to think very quickly and use devices that give us time for planning what we want to say and how to say it. Bygate (1987) calls this condition the *processing condition* and it refers to the fact that speech takes place under the pressure of time.

In addition to the processing condition, spoken language is influenced by the fact that it occurs between interlocutors. Though a written text also has a recipient, the writer seldom gets an immediate response from the reader. In speaking, the interlocutor may interrupt the speaker on the spot if the message is ambiguous, if he needs more background information, if he needs simpler language or a slower pace in order to understand, if he disagrees or if he simply wants to speak himself. The interlocutor therefore affects both the form and the content of the message. Bygate calls this condition the *reciprocity condition.*

The processing and the reciprocity conditions are generally assumed to account for the main differences between spoken and written language (Bygate 1987, Saleva 1997). In the following, each condition and its effects on speech will be treated separately.

3.2.1.1 The processing condition

As mentioned, the processing condition refers to the fact that the spoken language is planned and delivered simultaneously. Occasions where speech has been carefully planned and even written down beforehand, such as the lecture of a university professor, the sermon of a priest, the speech of the bride's father etc. are the exceptions that confirm the rule. This kind of oral presentation has a lot in common with written language and is rarely compared to the common use of oral language: small-talk to strangers, story- and joke-telling to peers, conversation with spouse and children, instructions to a colleague and explanations to a doctor or a car mechanic. In most situations where spoken language is used, planning of what one is to say next occurs as we speak. This affects the form of the spoken language in at least two different ways: on the one hand speakers use devices in order to *facilitate* production, on the other they use strategies to *compensate* for errors that occur due to the limited planning-time available.

3.2.1.1.1 Facilitation

There are four main ways in which a speaker can facilitate the production of speech under the pressure of time. Firstly, he may use *simplification*, both in syntax, morphology and vocabulary. *Syntactic simplification* is a characteristic trait of spoken language: utterances are connected by co-ordinating conjunctions or by no conjunction at all (*parataxis*) instead of by subordination (*hypertaxis*), which is more complex and therefore needs more planning time. In addition, sentences are shorter and less varied than in writing. As for *morphological simplification*, it is above all the verbal system that is affected. Speakers tend to use the present tense modified by the use of adverbs and time expressions in stead of using the past tense of the verb. In addition to the syntactic and inflectional simplification, speech is characterised by the simplification of vocabulary. The vocabulary is less varied, and speakers use more general and inaccurate expressions in speaking than in writing (Stenström 1994:1). The context and the common background knowledge of the interlocutors prevent misunderstandings from occurring too frequently. In addition, speakers tend to avoid complex noun-groups with many adjectives preceding the noun. Splitting up the noun-phrase makes speech less densely packed and easier for the listener to process.

A second tool used in facilitating speech is *ellipsis, which is* the omission of semantically redundant parts of the sentence, such as for example: "What?" instead of "What did you say?, "Sorry" instead of " I am sorry" etc. In a conversation there is always an amount of shared background knowledge. The use of ellipsis exploits this shared knowledge. Speaking in complete sentences and repeating known information would not be efficient communication and in fact, native speakers rarely do that.

A third way of gaining planning time as we speak is through the use of more or less conventional phrases, often called *formulaic expressions* such as "I don't know", "May I help you?", "Can I have a _____?", "I am sorry to _____",. "Would you like_____?" By using formulaic expressions, speakers do not need to plan each and every utterance from scratch, but can instead draw on expressions that are automated as chunks in the memory. This strategy frees capacity for the planning of other aspects of one's speech. Formulaic expressions may therefore be described as "islands of reliability" which give speakers time to plan their next utterances (Dechert 1983). Formulaic expressions may be considered in terms of a cline from very fixed combinations and idiomatic expressions ("Thank you!", "Nice to see you",) to half-fixed combinations ("May I have a_____?", "How nice of you to_____!"). In English the amount of formulaic expressions is estimated to be several thousand, depending on the definition of the concept, and it is probably similar for other languages as well (Saleva 1997:30, Pawley & Syder 1983:205). The use of formulaic expressions is a characteristic trait of early learner language and may be a useful tool in acquiring a second language.

A fourth way of facilitating speech is by using *time-creating devices*. As speech occurs in real-time and between interlocutors, there is always the danger that the interlocutor may be bored and decide to take the floor himself if the speaker spends too much time planning what to say next,. The speaker therefore uses devices that give him planning-time and at the same time signal that he does not intend to yield the floor yet. These time-creating devises include fillers, and hesitation devices such as right/ well/ okay/ you see/ kind of/ erm etc. (Hasselgren 1998).

Despite the use of facilitation devices to gain planning time, communicative problems do occasionally occur. Speakers need devices in to deal with such problems or to cope with a situation where the intentional meaning fails to get across. In addition to the facilitation devices, then, speakers need *compensation* devices.

3.2.1.1.2 Compensation

When spoken language appears to be less perfect than written language, this is partly due to the fact that mistakes occur and are corrected on the spot. In addition we often start an utterance only to abandon it half way through to start another, or we choose a word and replace it by another. Because of the lack of planning time, what we intend to say does not always get across. Bygate distinguishes between four kinds of compensation strategies: self-correction, false starts, repetition and rephrasing. *Self-correction* is when a speaker makes a mistake and corrects it himself immediately. It may occur on all linguistic levels. In speech, corrections are quite necessary, and do not disturb interaction unless they are excessive. *False starts* are also characteristic traits of oral

performance. The speaker starts an utterance, changes his mind, and leaves it unfinished to start in another direction. *Rephrasing* is also due to the limited time available. If the speaker assumes that his message did not get across successfully, he will choose a different path and rephrase what he just said. It is also quite common to *repeat* an utterance or part of an utterance when we speak. This repetition may include an expansion or a reduction of the original utterance.

When measuring the spoken proficiency of L2-learners, it is important to bear in mind that the compensation devices are common and useful tools in oral interaction between native speakers. They should not be regarded as mistakes or deficiencies, but as characteristic traits of the oral mode.

3.2.1.2 The reciprocity condition

The other main condition affecting the formal traits of the spoken language, as opposed to writing, is the *reciprocity condition* (Bygate 1987). This condition relates to the fact that (most) speech occurs between interlocutors. Not only do speakers have to plan the form and content of their utterances in real time, they also have to take into account the interlocutors' topical knowledge, their linguistic level, how much time they have to spend on the conversation etc. They have to be aware of their interlocutors' signals as to whether they find the subject interesting or whether they want to speak themselves and adapt their speech accordingly. The reciprocity condition therefore challenges our communicative sensitivity: native speakers are not equally good at this either. The reciprocity condition has a strong impact on the spoken language and has given rise to distinct methods of analysis and theoretical models such as interaction or discourse analysis.

3.3 Functional differences between writing and speaking

This far, the focus has been on formal differences between speech and writing. In addition, spoken and written language are normally used for different purposes, in other words, they fulfil distinct communicative *functions*.

The primary function of the written mode is the transmission of factual information (with the exception of fiction literature, poetry and personal letters). Brown & Yule (1983) call this the *transactional function* of language. Oral language is also used for information-transmission to some extent, but this is not its main purpose. Rather, speaking is for the most part a social activity: we speak to each other to establish and maintain social relationships (Vygotsky 1978, Brown & Yule 1983, Stenström 1994, Saleva 1997, Hasselgren 1998). Brown & Yule (1983) call this the

interactional function of language. Because this term may easily be confounded with the terms interaction and interactional skills, I have chosen to call this the *social function* of language. Each function will be treated separately in the following section, and the focus will mostly be on the spoken language.

3.3.1 The transactional function

A point was made above that the prime function of writing is transactional, while we speak to each other mainly for interactional purposes. As always when we make use of some kind of classification, oversimplification seems inevitable: obviously, we do transmit information in speaking too. In fact, in one and the same conversation some turns are transactional while others are purely social. In a conversation the opening and closing, or the warming-up and winding-up talk often contains very little factual information, while the factual content of the message is transmitted in between. Transactional spoken language is "frequently concerned to get things done in a real world" (Brown & Yule 1987): people give messages about what they want their peers, colleagues, hair-dresser, doctor or dentist to do for them. Similarly, we may try to convince our interlocutors about our point of views in a discussion. In such situations getting the message across successfully is of major importance. Other examples of transactional language use in speaking are "telling a story or a joke, giving route or other instructions, reporting an accident to the police, or taking a stand in a debate" (Saleva 1997: 34).

Whether we speak to transmit information or simply to socialise affects the form of the spoken language. As the purpose of transactional language use is the transmission of factual information, it is of great importance that the message gets across in a clear and unambiguous way. The speaker therefore has to use more specific vocabulary than he would if the purpose were simply to socialise. He has to structure his turns in a logical manner, presenting the subject of the message, checking the interlocutor's background knowledge to tailor his speech to the listener, making a request for understanding as he speaks and after the message has been delivered. When the function is transactional, speech tends to be clearer and more precise and the turns longer than when the function is social. In cases where the intended message fails to get across, the interlocutor will normally signal that he does not understand, and the speaker will repeat his message. The focus of transactional speech is the message itself. Consequently transactional speech may be labelled *message-oriented.*

It is important to bear in mind when we assess the spoken language of non-native speakers that the ability to structure long turns of factual information in a clear, well-organised and unambiguous way varies between native speakers as well.

3.3.2 The social function of speech

While the purpose of transactional language use is to get the message across, the main purpose of social speech is to establish and maintain social relationships. So while transactional language may be referred to as message-oriented, social language is *listener-oriented*: we talk to our neighbours, peers and colleagues, not only when we have important information we want them to know of, but rather to show a friendly and pleasant attitude. Even though social speech does have a content, the importance of getting the message across is less important than making the interlocutor feel at ease. The speaker may therefore allow himself to be rather unspecific and general in his formulations. Of the same reason, if the listener disagrees he will normally not let the speaker know. Nor will he signal lack of understanding by asking for repetition. Social language is characterised by short turns and frequent shifts of topics. These characteristics of the social function of speech are particularly obvious in the phatic function of language. In this kind of language the topics are general and uncontroversial and of a kind that does not encourage disagreement (the weather, the speakers' health or the immediate surroundings). It is often used in the opening and closing of conversations as well as in conversations with people we do not know very well.

In Bachman's general model of CC (1990) both the transactional and the social functions of language are categorised as part of a language user's pragmatic competence, under the sub-component of illocutionary competence and *manipulative functions*. According to Bachman, the main purpose of both functions is to affect the world around us, either by getting things done or by establishing and maintaining social relationships. Bachman and Palmer (1996) use a similar categorisation, but the term illocutionary competence is replaced by the term functional knowledge in the later model.

The main purpose of this chapter has been to give an overview of the spoken language, with focus on its form as well as on its distinct functions. Differences between writing and speaking were discussed, and it was argued that these differences are due to the influence of two conditions affecting speech: i.e. the processing and the reciprocity conditions. The point was made that the most common kind of oral language is conversation or interaction. I have shown that interaction may fulfil different purposes: we may speak to transmit information (transactional language use) or with the purpose of socialising (interactional or social language). The different

functions of speech affect the form of the language used. By reference to Bygate's model, the distinct sub-skills of speaking were presented.

The main subject of this chapter was speaking in a first language context. In relation to the measurement of non-native speakers' language, there are two important points to bear in mind. Firstly, it is important to acknowledge that spoken language is not written language spoken aloud. It is characterised by such traits as simplifications, ellipsis, hesitation, self-correction, repetition and rephrasing. Hence, compared to a written text, spoken language appears to be less perfect even in a native speaker context. For a fair assessment of L2-speech, then, it is of crucial important that the candidates' performances be compared to native speakers' oral performances and not to some norm for the written language, a point also made by Saleva 1997:23.

Another point made in this chapter is that some kinds of oral language use are difficult even for native speakers. Some native speakers have problems with turn-taking, while some find transactional and message-oriented speech difficult. It is worthy of discussion whether abilities which are difficult even for native speakers should be measured in a foreign language test at all.

CHAPTER 4: BASIC MEASUREMENT CONCEPTS

In order to discuss practice and problems in relation to the measurement and scoring of speech, a test-theoretical framework is required. The main purpose of Chapter 4 is to introduce a set of general concepts in relation to measurement. Oral testing, more specifically, is treated in three parts: characteristics of oral tests and frequently used test methods are discussed in Chapter 5, rating procedures, rating criteria and rating scales are treated in Chapter 6, while the rater-variable and rater effect on test-scores are the focus of Chapter seven.

In everyday life, as well as in the literature of the field, "measurement", "test" and "evaluation" are used more or less synonymously. Bachman (1990:18) stresses the importance of a clear definition of these concepts. Even though they may refer to the same activity and to some extent overlap, they do have their distinct characteristics. *Measurement* in the social sciences is defined in Bachman (1990:18) as "the process of quantifying the characteristics of persons according to explicit procedures and rules". The process of assigning numbers to a person's characteristics is a crucial attribute of measurement. It is not until this is done that we may speak of true measurement. Tests where a candidate's characteristics are reported verbally (*A, B, C…* or *poor, good, excellent*) can therefore not be characterised as measurements. However, quantification is not in itself a sufficient property of measurement. As the definition states, the assigning of numbers has to be done according to "explicit procedures and rules". In subjectively scored tests this means developing rating criteria and rating scales to ensure that raters set their scores according to a common procedures.

A *test* is a measurement instrument, and the demand of quantification therefore applies to tests as well. What distinguishes a test from measurements in general, is that it is designed to obtain a specific sample of behaviour. Carroll (1968:46) defines a test as: "[…] a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual". While measurement often involves the collection of a set of different samples of a person's characteristics over a period of time, a test aims at eliciting a precise sample which makes it possible to make inferences about the test takers' abilities in other situations. In other words, a measurement may be based on several tests.

Evaluation is defined as "the systematic gathering of information for the purpose of making decisions" (Weiss 1972). The data used in evaluation may be quantitative and collected through the use of tests, but it may just as well be qualitative and based on the teacher's verbal description of a pupil. Both measurement and tests are systematic gathering of information and

therefore share common characteristics with evaluation. But tests and measurements need not be used for the purpose of making decisions. They may be used for other purposes as well: students may find tests motivating in the course of instruction, or teachers can use tests and measurement in order to gain insight into the students' proficiency levels. Bachman stresses the importance of distinguishing the information-providing function of measurement on the one hand, and the decision-making function of evaluation on the other. In this thesis I am concentrating on tests as measurement instruments, while evaluation is left aside.

4.1 Different types of language tests

When investigating the quality of a given test, we need to know what kind of test we are dealing with: what is the purpose of the test? What are the test-results being used for? Are important decisions being made on the basis of the scores? Is the test part of an educational program, or is it independent of instruction? What test-methods and test-tasks are included in the test, and how is it scored? All of these questions affect the demands we make as to the qualities of the test. While one test may only have a limited significance internally in a teaching course, another may me high-stake and decide whether a person is allowed entrance to higher education or is permitted to practice his or her profession in a new home country.

In the literature, reference is often made to distinct kinds of tests without a specification of the criteria by which the tests are grouped. Bachman (1990) gives a useful categorisation of language tests according to five criteria: 1) intended use, 2) content, 3) frame of reference, 4) scoring procedure and 5) testing method. Bachman's categorisation will be used here in the discussion of different tests.

4.1.1 Intended use

Tests may be categorised according to their purpose or intended use. Bachman distinguishes between tests used for collecting information for research on the one hand, and tests used in educational settings, on the other. In SLA research, tests can be used to gather information about the effect of different pedagogical methods, about differences between learner groups or about the very nature of language and language acquisition. The kind of test one chooses to use depends closely on the research questions of the study.

In educational settings, tests can be classified according to the kind of decisions that will be based upon them: when decisions are to be made regarding a person's qualifications for entering an instructional program, the test is referred to as a *selection, entrance* or *readiness* test. Once in the instructional program, *diagnostic* and *placement* tests are used in order to offer the student

instruction at a level corresponding to his or her proficiency. As a means to find out about students' progress during the instruction or success after instruction a test may also be used: such tests are referred to as *achievement, attainment* or *mastery* tests.

In addition, tests may be used for making decisions regarding whether a person may practice a profession for which the studies have been undertaken in a different country. This is the case for medical personnel in Norwegian society. In this case, the test is used for *selection* or *entrance* but instead of entrance to universities or other educational settings, it concerns entrance to the labour market.

4.1.2 Content

A commonly referred dichotomy where different language tests are concerned, is the one of proficiency versus achievement tests. This opposition refers to the content on which the tests are based: an *achievement test* bases its content on an explicit syllabus. The syllabus states the aims of instruction in relation to both factual knowledge and linguistic skills, and the test-scores relate to the candidates' mastery of the syllabus.

Proficiency tests on the other hand base their content on a theory of language proficiency. They are often independent of instruction, so that a person may take the test to learn about his or her mastery of a given language without having followed a teaching course first. Whether the content of a given achievement or proficiency test actually measure the same thing, depends on whether the syllabus and the proficiency test are based on the same underlying theory of language proficiency (Bachman 1990:71).

A special kind of proficiency test is the *language aptitude test,* which were used to a great extent in the USA between 1925 and 1960 (Spolsky 1995:117). Language aptitude tests are theory-based, but rather than a theory of language proficiency they are based on a theory of language aptitude which describes the relationship between the aptitude for learning language and other cognitive, motivational and personality factors. Such tests are only used to a very limited degree today.

4.1.3 Frame of reference

Tests may also be classified according to the frame of reference within which their scores are interpreted. The test literature distinguishes between norm-referenced and criterion-referenced measurement.

Norm-referenced (NR) measurement is "an approach to measurement in which an individual performance is evaluated against the range of performances typical of a population of similar

individuals" (McNamara 2000: 135). In other words, the score of each test-taker is compared to the scores of other test takers. These "other test-takers" may be a group of individuals considered to be representative of the test-taker population, and who take the test in order to set a norm for comparison for future test-takers. An assumption underlying NR- tests is that test-scores will be normally distributed in the same way as physical attributes such as height or weight. Test takers may use a norm-setting group in order to describe the normal distribution of scores on the test. Yet, in much NR- measurement, test scores are not compared to the performance of some external group, but rather to the performance of other test-takers within the same group. This is the case when runners in a race are ranked as first, second and third or when pupils' written essays are compared and ranked in accordance with each other. Often in NR-measurement a certain percentage of the pupils are to be given a certain grade irrespective of the quality of their performance.

In *criterion-referenced (CR) measurement*, on the other hand, performances are compared not to other performances, but to "one or more descriptors of minimally adequate performance at a given level" (McNamara 2000: 132). The performance of each test-taker is compared to verbal descriptors of some rating scale and given a score without reference to other test-takers of the same group or of a norm-setting group. Central to this kind of assessment is the development of adequate descriptors for each level of the grade scale. An example of CR- measurement, is the evaluation of pupils in a school class where all pupils fulfilling the demands for obtaining an A, are granted that score, irrespective of the number of A's in that class or in a normal distribution of similar pupils.

This thesis focuses on the use of rating scales and raters' ability to place test-takers on the scale in a reliable and valid manner. The test under scrutiny, *Språkprøven i norsk for voksne innvandrere*, is criterion-referenced: it makes use of explicit rating scales and trained raters (Pedersen 1997). Norm-referenced assessment, and ranking of candidates, is therefore irrelevant for the research questions raised here.

4.1.4 Scoring procedure

Tests are sometimes referred to as subjective or objective, but as Pilliner (1968) points out, this is misleading. The subjective/ objective dichotomy does not really refer to the tests per se, but to the scoring procedures involved. The construction of both kinds of tests involves subjective decisions: in both test types, test-constructors choose which texts to use for reading

comprehension, dictation or cloze, which issues to be raised in the free composition or oral interview, what distractors to use in the multiple-choice task etc. The difference between an objective and a subjective test is that while an *objective test* can be scored mechanically without the use of human raters, *subjective tests* can only be scored in a proper way by using human evaluation. A typical example of an objective test is the multiple-choice test. Oral interviews and free written compositions, on the other hand, are associated with subjective scoring procedures. It is repeatedly stressed in the literature that these skills cannot be scored in a satisfactory way without involving the use of human raters (McNamara 1996, 2000).

4.1.5 Test methods

A final distinguishing feature of tests, is the test method used. Considering the fact that there is a multitude of test methods, and that one and the same test may include a set of test methods, it is hard to draw a sharp line between different tests according to this feature alone.

Test methods may be classified according to some commonly referred to dichotomies, however: direct versus indirect, discrete point versus integrative, authentic versus inauthentic or paper-and-pencil versus performance.

A *direct* test method aims to measure a trait by the shortest route, that is, without passing through other traits or abilities. A direct test of oral ability will therefore require the test-taker to speak, and similarly a direct test method of written production would require the test-taker to produce a written text. Direct test methods are taken to hold high face validity, and be motivating for test takers. The obvious drawback is the time and costs involved: both eliciting and scoring of these skills directly takes time, and it requires personnel, both as examiners and as raters.

Despite the obvious advantages of direct test methods, *indirect* measures are sometimes preferred in order to diminish the expenses involved. This is particularly the case in large scale testing where the measurement of written and spoken language would otherwise be beyond the scope of economic realism. In an indirect test method, the trait one wishes to measure is elicited through the measurement of other traits, and the validity of the indirect measure is established through concurrent validity between the indirect method and a direct method of the same trait. An example of an indirect measure of written production is the one used by the DIALANG project. Since the tests of DIALANG are data-based and scored automatically and objectively by the computer, a direct measure of written and spoken language is impossible. The test does however include the measurement of indirect writing, where test takers are asked to arrange paragraphs of a text, to fill in missing words in a text etc. The validity of the test would have to

be established by an a posterior investigation of the correlation between the scores on this test with the scores of a more direct measure of written production.

The other commonly referred to dichotomy for test method classification is that of integrative versus discrete point tests. A *discrete point* test typically contains a set of items focusing on one trait of grammar or vocabulary at a time. It is useful in diagnosing the test-taker's mastery, or lack of such, of specific points of the language. The test-items are however context-free, and the test does not yield information about how the test-taker would perform in actual communication where several language components interact.

Its opposite is *integrative* test methods. Such tests measure several language components at a time, and they focus on the integrated knowledge necessary for the test taker to perform adequately in context. Examples of integrated test methods are cloze tests or oral interview. In both cases the test-taker has to use pragmatic as well as organisational competencies (Bachman 1990). It is also sometimes referred to as integrative when one and the same test measures several skills simultaneously, such as listening and speaking, or reading and writing. A listening comprehension test, where the test taker is required to answer in writing includes the measurement of writing abilities as well as that of listening ability.

Authentic versus inauthentic are other labels used when referring to test methods. An *authentic* test method is one that resembles situations the test-taker may encounter in real-life. As the purpose of a test score is to generalise to some situation outside the test, authentic test methods are taken to yield more valid scores: it is assumed to facilitate interpretation and generalisation of the results to non-test situations when the test method reflects real life performance. A typical authentic test method is a conversation between peers when this conversation includes an information gap: making one candidate find out as much as possible about his or her co-candidate, is a task that resembles real life. Asking someone questions, giving responses, asking for clarifications, and thereafter talking about yourself, are authentic, real, life-like tasks.

An *inauthentic* test method, on the other hand, is one that the candidates will meet exclusively in a test-situation: filling in gaps in a cloze-test, or choosing between alternative forms of some of the words in a multiple choice test of reading comprehension, are examples of inauthentic test methods. However, inauthentic test methods may elicit language in a valid and reliable way and even have positive washback effect on teaching. In recent test literature the assumed inferiority of inauthentic methods, as compared to authentic methods, is challenged: a test situation is not a tea party, but it is still a part of life, hence it does not have to resemble a tea party to yield valid scores (Spolsky 1985)

Finally, a line is drawn between paper-and-pencil and performance testing. A *paper-and-pencil* test is associated with a traditional test format, where test takers fill in an answer sheet. The items are often discrete-point, and the test may in most cases be corrected automatically and objectively.

A *performance* test measures the candidates' ability to perform in a language situation. Such test methods typically require the candidate to produce a rather large amount of language, in writing or in speaking. Hence they cannot be scored without the evaluation of human raters. A critical point in performance testing, then, is the development of valid rating scales and the training of raters.

4.2 Qualities of tests

A chief occupation for language test constructors and researchers is to develop high quality tests, as well as to investigate the quality of already existing tests. But what distinguishes a good language test from a poor one? A good test measures a given skill or ability in a consistent and fair way so that two candidates who perform equally well are given the same score (reliability). In addition, a good test measures whatever the test constructors assumes it to measure, and not other related skills or abilities (validity). Moreover, a good test is practical and economic in use (practicability). And finally, it has a positive effect on language teaching, so that it stimulates activities that are actually known to encourage acquisition (positive washback effect). While practicability and washback effect are relatively unproblematic concepts, reliability and validity are complex concepts which need to be discussed in more detail.

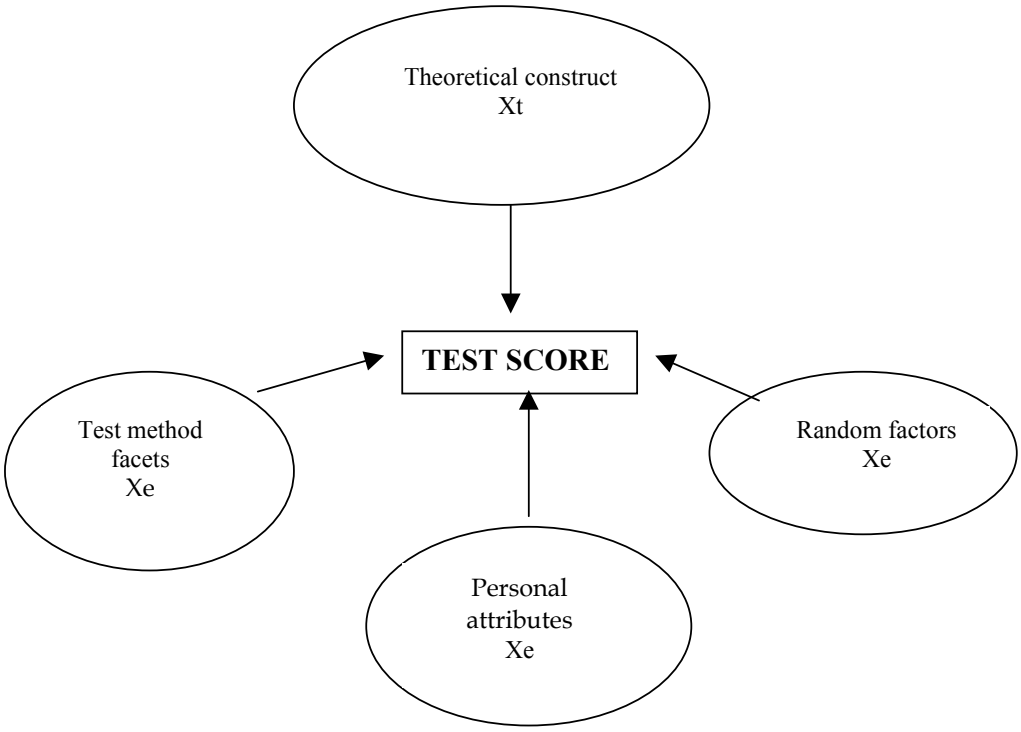
4.2.1 Reliability

Reliability is a basic sign of quality of a test score. It relates to the degree of stability of a measurement instrument across raters and test administrations.

> Whenever a test is administered, the test user would like some assurance that the results could be replicated if the same individuals were tested again under similar circumstances. This desired consistency (or reproducibility) of test scores is called **reliability** (Crocker and Algina 1986:105).

If there had been a total correlation between a test score and the construct of the test, i.e. whatever the test sets out to measure, the candidate would be given identical scores across test administrations. This is hardly ever the case in language testing, due to the fact that a test score will always be affected by factors other than the ability in question. This acknowledgement is the basic underlying assumption of the true-score model as presented below:

Figure 5 Factors that affect language test scores Bachman (1990: 165).



A test score is affected by whatever one wishes the test to measure, the theoretical construct or the *true score*, Xt, as well as by other variables, *error score,* Xe. The purpose of reliability investigations is to shed light on these potential sources of measurement error. Bachman (1990) distinguishes between three main sources of error score:

- Test method facets (test tasks, rating scales, raters)
- Personal attributes (sex, age, ethnic background, prior knowledge of content area, cognitive style etc.)
- Random factors (the candidate's mental and emotional state, differences in test administration from one place/ time to another etc.)

The degree of reliability in measurement depends on the test constructors success in isolating the ability in question and minimise the effect of other factors affecting the score. This requires knowledge about the theoretical construct, communicative language ability in our case, as well as the potential sources of measurement error.

For subjectively scored tests, such as direct tests of oral and written production, variation due to the rater evaluation is a potential source of measurement error. Rater related reliability is of two kinds: *inter-rater reliability* concerns the agreement about the scores awarded by different raters and answers questions such as: "Is the same candidate assessed identically by different raters"? *Intra-rater reliability*, on the other hand, has to do with the internal consistency of each rater. A relevant question for this kind of reliability would therefore be: "Are different candidates scored consistently by one and the same rater"?

In Classical Test Theory (CTT), which is the framework of the true score model, it is the aim to reduce variability and thereby obtain as high a reliability estimate as possible, that is an estimate as close as possible to 1.00 (Crocker and Algina 1986). This view is challenged in Item Response Theory (IRT) and particularly with the introduction of multi-faceted Rasch analysis (MFR, Linacre 1989). Within this framework a certain degree of variation between raters is seen as necessary in assuring test validity: communicative language ability in speaking and writing is multi-componential and so intricate that a complete agreement between raters would jeopardise a true evaluation of all aspects of the skill, it is argued. Some disagreement is considered natural and inevitable, and even necessary for a valid assessment of all aspects of the skill (Huot 1990, Alderson 1991b:64, McNamara 1996:232). Another reason why the importance of rater-agreement about the scores is diminished in this approach is that MFR and the data program FACETS may actually take differences between raters into account and correct for such variability. Still it is important that each rater be internally consistent[3].

4.2.2 Validity

A common definition of validity is the degree to which a test measures what it is meant to measure (Hughes 1989:22). This definition is however imprecise: validity is a quality of test scores and not of tests per se. It is repeatedly stressed in the literature that no test can be valid for all occasions and for all uses (Messick 1975, Bachman 1990). Hence, when examining the validity of test scores, we need to take into account the purpose and use of the test. One and the same test may yield valid scores as a basis for a teacher's assessment of his or her students, but be invalid as a basis for important decisions such as university entrance. In more recent articles, Messick argues that validity refers to "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (Messick 1989:13). Relating validity not only to test scores, but to the interpretations and actions

[3] For arguments why I still choose to focus on inter-rater reliability in the present study, se Section 8.1.

based on the scores is controversial. It asserts that test constructors could be held responsible not only for their tests, but for possible misuses of their tests as well (Alderson and Banerjee 2001:79). Obviously, test constructors cannot prevent their tests or test results from being misused. Yet, precise definitions of the construct as well as of its scale descriptors would probably render the test less prone to misuse.

Early test literature typically referred to different kinds of validity. In an overview of the concept Angoff (1971) cites as many as 16 different kinds of validity mentioned in the literature. Some of the most frequently mentioned are content validity, face validity, concurrent validity, predictive validity and construct validity. However, modern scholars have largely abandoned this divided notion in favour of a unified concept.

> Validity […] is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores (American Psychological Association 1985).

In a series of articles in the 1980s and 90s, Messick argued for a unified concept of validity with construct validity as the umbrella under which the other types are subsumed. This view has reached consensus in the field, and it is now common to refer to construct validity as a "multifaceted but unified and overarching concept which can be researched from a number of different perspectives" (Alderson and Banerjee 2001:79). In this view, content, predictive and concurrent validities are not different types of validity but different aspects of construct validity.

CHAPTER 5. MEASURING SPEECH

This chapter begins with a discussion of the status of oral tests in three different directions or periods in language testing during the last century. Its role in the communicative paradigm of today is also outlined. Thereafter follows a discussion of characteristics of oral tests building on the classification features introduced in Section 4.1 above, before the relation between the test construct and test methods is outlined. The main part of the chapter consists of a presentation of some commonly used test methods for eliciting speech, and their advantages, as well as their disadvantages, are discussed.

5.1 The status of oral tests

Oral tests held a varying status during the last century, mainly due to problems in ensuring their reliability. Since oral performance has to be assessed by human raters, the value of oral tests is closely connected to whether or not society has confidence in the raters' ability to judge consistently. In this section I will discuss the status of oral tests in three different periods of the history of language testing (Spolsky 1976, 1995). While these periods replaced each other successively in the US, they exist side by side in other parts of the world. Hence, for language testing outside the US, *approaches* or *directions* would probably be more appropriate labels than periods.

5.1.1 The pre-scientific period/ direction

The pre-scientific period or approach to language testing is characterised by its "lack of concern for statistical matters, or for such notions as objectivity and reliability" (Spolsky 1976: 11). The typical test of this approach is a subjectively scored test of free written or spoken production. The pre-scientific approach is based on a complete reliance on the rater's ability to grant adequate scores, using his or her experience as a teacher and rater alone. Raters are not requested to explain why they set a particular score, and rating criteria and rater training are not common procedures. There is a naïve reliance in tests and raters, and questions regarding what the test really measures, or whether it is fair and reliable, are hardly raised.

In the US, the pre-scientific period came to an end as early as by the 1920s. In England it held its ground till the 1970s, while it is still the prevailing trend in many countries today. In Norway, quite a lot of language testing in schools and universities has to be characterised as pre-scientific: teachers set the scores, though often together with an external rater, yet rating criteria

and rater training are not common practice. The reliability of raters is only to a very limited degree subject to discussion. However, during the last few years, there has been research drawing attention to the potential unreliability of untrained raters (Berge 1996, Raaheim 2000, 2002, Carlsen 2003) or other test-related issues (Hellekjær in progress). In addition, a new reform of higher education, "Kvalitetsreformen 2003" focuses on different aspects of the quality of higher education, including the routines for student assessment. The introduction of a new grade-scale as part of the reform has led to some discussion about the scoring procedures and their potential lack of reliability (Carlsen 2002). Another exception from the pre-scientific tradition in Norway, is the work of Norsk språktest (NST), which has been developing language tests in accordance with international standards of professional test construction and evaluation since the 1980s. Norsk språktest is also responsible for publications and conference lectures, which are central in the emerging of a new scientific discipline (Andersen 1995a, 1995b, 1999, Moe 2002, 2003a, 2003b, Halvorsen 2002, 2003, Dregelid 2002, Salomonsen 2002). There is therefore reason to believe that the interest in and consciousness about test related issues are growing.

5.1.2 The psychometric-structuralist period/ direction.

In the US, the reliability of subjectively scored tests was questioned as early as towards the end of the nineteenth century (Edgeworth 1890). Those who were submitted to the tests, as well as social institutions that were to interpret the test results, started to regard the existing tests with scepticism, and there was a growing plea for objective, reliable and fair tests. At the beginning of the twentieth century, a new direction, which Spolsky has named the psychometric-structuralist period, came into view. This direction is described as a fusion of linguistic structuralism and psychometrics[4].

The new approach emerging in the US during the 1920s and 30s focused on the measurement of discrete items separately based on the structuralist view of language as a set of finite structures. In order to reduce the uncertainty and unreliability of earlier tests, methods that could be scored objectively were preferred. The typical test-method of this direction was the multiple-choice format. Indeed, this direction was characterised by a conviction that it was possible to measure language ability in an exact and objective manner. The traditional measurement of written and spoken production did not satisfy the demands of objectivity and

[4] At the turn of the century psychologists started the development of large-scale intelligent-tests which could be scored objectively. This development gained speed with World War 1 and the IQ-testing of all military recrutes. The experience from the large-scale IQ-testing was later exploited in the measurement of language ability.

reliability postulated in this direction and was therefore largely avoided (Spolsky 1995:30). The consequences of this approach are expressed in Spolsky (1995:54):

> The quantifiable results provided by the mechanistic scoring of short true-false or multiple-choice questions, and the opportunity that large numbers of marks afforded of replacing judgements of individual performances by statistical norms, gave every appearance of solving the problem of reliability. Whatever it was that was being measured, at least it was measured consistently" (Spolsky 1995:54).

However, towards the end of this period, one gradually came to realise that *what* is measured is, and indeed should be, as important as *how* it is measured. The potential unreliability of oral tests is a challenge that has to be dealt with in other ways than by excluding the measurement of oral and written production altogether.

The importance of being able to speak a foreign language became pronounced with the participation of the US in World War ΠI: diplomats and other civil servants needed to be able to speak a foreign language in order to perform their overseas duties. The Foreign Service Institute (FSI) was commissioned to improve the speaking skills of overseas personnel, and in the 1950s the FSI started the job of developing a test that measured spoken production without violating the demands of reliability so ear-splittingly pronounced in the psychometric-structuralist period (Fulcher 1997:76). In order to reduce the subjectivity involved in the scoring of speaking, two procedures were emphasised: the development of rating scales with explicit criteria upon which raters should base their scores, on the one hand, and the training of raters in interpreting and implementing the scales, on the other (Spolsky 1995:177). These efforts to ensure a fair and reliable assessment of speaking and writing is perhaps the most valuable result of the period.

5.1.3 Psycholinguist-sociolinguist period/ direction.

The psycholinguist-sociolinguist period of language testing differs from the previous period by its conception of what language is, how it is learned, and how it should be measured. While the structuralists described language as a finite set of structures which could be taught, learned and assessed separately, this period focused on the integrative and creative character of language. This is evident in Carroll's plea for integrative items in language tests:

> […] I recommend tests in which there is less attention paid to specific structure-points or lexicon than to the total communicative effect of an utterance…Indeed, this "integrative" approach has several advantages over the "discrete structure point" approach (Carroll 1961:37-8).

Cloze-test and dictation were the recommended test methods assumed to satisfy the criterion of integrative testing.

This period saw a shift in focus from reliability to validity as the prime virtue of language tests: what to measure, should be more important than how it is measured, consequently language production should be measured despite the problems in obtaining the same degree of reliability as with objectively scored tests.

5.1.4 Communicative language testing

The main ideas of the psycholinguistic-sociolinguists' approach are maintained in what is referred to as the communicative trend in language testing of today, yet after the 1980s the importance of measuring a candidate's ability to communicate in real situations has been argued more vigorously. In addition to the practical-pedagogical appeal for tests mirroring the second language (SL) didactics's focus on communication, the theoretical foundation of such tests was strengthened during the 1980s and 90s (see Chapter 2).

Communicative language tests aim at measuring not the candidates' knowledge of language, but their ability to use their language skills in actual communication. As Morrows states:

> Knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language (Morrow 1979: 145).

According to Weir (1990:30-31) a communicative language test has the following characteristics: it emphasises *interaction* either between candidates or between the candidate and the examiner. It is relatively *unpredictable* as it gives the candidates freedom to elaborate their speech and take extensive control of the content and form of their production. It strives at being *authentic* in the choice of texts as well as in the test methods used. Moreover, it is *integrative* in that it opens up for a simultaneous assessment of several language skills. It prefers the use of *direct* test methods over indirect. And finally, it is primarily interested in measuring *language production* over the measurement of the receptive language skills, which were the focus of the psychometric-structuralist approach. The test which best meets with the demands outlined above, is the direct test of speaking (Underhill 1987, Alderson and Hughes 1981:60, Weir 1990).

5.2 Characteristics of oral tests

What characterises a test of oral ability, and in what ways is such a test different from other language tests? The obvious answer is of course that an oral test yields information about a

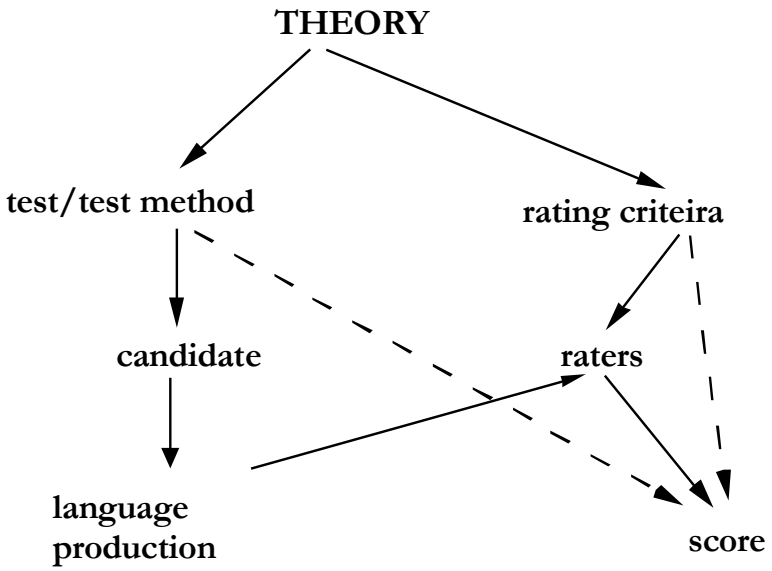
candidate's ability to speak. And because the skill of speaking is different from the skill of writing, reading or listening, the underlying construct of an oral test is distinct from tests of other language skills. Yet, it has other characteristics as well: normally, the candidates are expected to produce speech of some kind, either in a presentation, an interview with the examiner or in a peer conversation. The performance is evaluated by one, or several, human raters. In most cases the scoring procedure is standardised and raters base their scores upon rating criteria verbalised in a rating scale.

The five classification features mentioned in the previous chapter were intended use, content, frame of reference, test methods and scoring procedures. The first three features do not distinguish oral tests from tests of other language skills: an oral test may be used for different purposes: as an entrance, a diagnostic, as well as an achievement test. It may be theory-based or syllabus-based, or in other words a proficiency or an achievement test. As to the frame of reference it may be norm-referenced as well as criterion-referenced. The two remaining features, test methods and scoring procedures, are the ones that distinguish an oral test from other language tests and the ones worthy of discussion in a chapter on the measurement of speech. The scoring procedures of oral tests are at the very core of this project. It will therefore be treated in detail in the next two chapters dedicated to this: Chapter 6 focuses on the rating scales and rating criteria for oral tests, Chapter 7 on the effect of the rater variable on test scores.

5.2.1 Test methods for measuring speech

Test methods are closely related to whatever one wishes to measure with a particular test. Different concepts of what language ability is will therefore inevitably affect the choice of test method. An oral test in the structuralist-behaviourist paradigm would focus on specific structures of the sound system separately. The communicative paradigm, on the other hand, would use test methods resembling real life tasks, where the candidates are involved in actual communication and where integrated skills are applied. The relation between the underlying construct and test methods is graphically presented in Figure 6 below.

Figure 6 The relation between test construct and test methods.



Different test methods are suited for eliciting different kinds or aspects of speech. To measure what one wishes to measure, then, it is necessary to define the construct and also to know which methods are most suitable for eliciting that particular construct. In the first two chapters of this thesis, two possible constructs for oral tests were described: Bachman's model of communicative language ability on the one hand, and Bygate's model of speaking, on the other. When measuring spoken language, it is a crucial decision whether to assess a general language ability, which could equally well be assessed through a test of writing or, on the other hand, whether one should, rather, focus on aspects of language which are specific for the spoken mode and which could hardly be assessed through anything but an oral test, such as pronunciation and fluency as well as interaction skills. Another problematic issue when defining the construct is the treatment of skills that are difficult even for native speakers, such as long transactional turns and presentation of factual information in a coherent manner. The questions test constructors should raise, then, is whether it is fair to include such aspects in a spoken test for non-native speakers. The answers to these questions depend of course on the purpose of the test, the underlying theory of language and the people for whom it is constructed. If the test is used as a university-entrance test, it is reasonable to request skills that not all native speakers would perform equally well on. If, on the other hand, it is a test at an intermediate level for non-academic purposes, assessing the ability to present factual information as in a prepared speech would be more problematic from an ethical point of view. However, knowledge about language ability and definitions of the construct are

not sufficient for constructing an oral test, which yields reliable and valid scores. In addition, it is necessary to be aware of the effect of the test method on test scores.

Different methods elicit different kinds of speech (Underhill 1987, Weir 1990, Fulcher 1997). A conversation between two or more candidates elicits interactional skills, including turn-taking, question formations, requests for information and repetition, while these skill are hard to elicit through a controlled interview. If interaction and the ability to ask for clarifications are among the rating criteria, it is of course fundamental that the method used is suited to eliciting these skills.

Secondly, the test method is a potential source of error score. One and the same candidate may perform quite differently on a role-play, a read aloud or an oral presentation. It is therefore important that test takers be familiar with the test method used in the test and also that the test includes different test methods. It is of paramount importance that test constructors be aware of the fact that the choice of one test method over another affects test scores. There are problems and drawbacks connected to all methods. Test constructors need to be aware of these and bear them in mind when scoring the candidates, as well as when interpreting the results.

Finally, some test methods used in language tests elicit other skills and abilities than language ability in isolation. A read-aloud test measures the candidate's ability to read aloud, and not only his or her pronunciation. A role-play elicits the candidate's imagination and acting skills, and an oral presentation his or her factual knowledge and academic skills. This may jeopardise the construct validity of scores, unless, of course, these other skills are defined as part of the construct of the test. Some of the most commonly used methods for eliciting the spoken language are referred to below.

5.2.1.1 Controlled interview

Of all oral test methods, the controlled interview is probably the most frequently used. In the interview the examiner has a more or less fixed list of written or memorised questions to ask or topics to raise. The interview is structured, and the examiner controls the content. Yet test takers are given the freedom to elaborate their answers in their own manner, and thereby influence the course of the interview. A successful interview turns into a genuine conversation between the examiner and the examinee, rather than being a mere sequence of questions and answers. If the former happens, the interview may be a reasonably authentic test method.

Since the topics and questions are fixed, different test takers get approximately the same task across examiners. This is an advantage for comparison of performances across candidates, which in turn should affect test reliability positively. If the interview develops into a real, life-like conversation, it has the potential of eliciting not only the candidate's motor perceptive skills but

his interactional skills as well. If related to Bachman's model of communicative language ability, an interview which works well measures the organisational competence and the pragmatic competence, as well as the test taker's strategic competence.

The interview has been criticised for not being suitable for eliciting crucial parts of actual language use. As Weir emphasises in his critique of this technique: "In interviews it is difficult to replicate all the features of real life communication such as reciprocity, motivation, purpose and role appropriacy" (Weir 1990:76). The interview often fails to develop into a genuine discussion, and the validity of a mere question and answer sequence is worthy of discussion. The fact that the candidate is talking to a person who represents authority will also affect the conversation: at least for candidates from hierarchical cultures, it may feel unnatural to ask questions or ask for repetitions from someone you consider an authority such as the examiner.

5.2.1.2 Candidate-candidate discussion

Some of the problems related to the interview test may be overcome by replacing the examiner by an examinee. In a conversation between two or more peer candidates, test constructors may gain at least two things: firstly, this test method has been proved to reduce test anxiousness. Test-takers report that they feel more at ease when talking to a peer candidate than to an authority native speaking examiner. In fact these peer conversations often become very similar to real life conversations, and candidates seem even to forget that they are in a test situation, and start to focus more on the content than on the formal aspects of their speech. Secondly, in a conversation between candidates it is easier to make candidates produce real, life-like language including questions and requests for more information or repetition when they fail to understand. It may be a very good test of interaction skills (Fulcher 1996b, Weir 1990:78).

One obvious problem in relation to this technique is how to match the candidates so that both of them get the opportunity to perform their best. We lack empirical evidence about the effect of matching uneven candidates on test scores, but it is reasonable to assume that it does have an effect. Different personality types may also influence the peer conversation, yet how personality types such as introvert versus extrovert, shy versus talkative, confident versus nervous etc. affect the results of the test it is not well documented. For this method to function, it depends on the examiner's ability to notice such things and make sure that both candidates get a chance to produce enough language for the raters to score, irrespective of their language proficiency or personality type. *Språkprøven i norsk for voksne innvandrer* has in part solved this problem by having several parts of the oral test, some of which are candidate-candidate discussion, and some which are individual tasks (Norsk språktest 2003)

5.2.1.3 Oral presentation

In an oral presentation the candidate is given a topic on which to prepare a short talk either days in advance or shortly before the test. The candidate may use notes but is encouraged not to read from the notes. The oral presentation is claimed to be "an authentic and communicative activity both for professional and academic purposes" (Underhill 1987:47). The method is authentic in the way that we sometimes have to present different topics as part of a normal conversation or in an educational or professional setting. It elicits all parts of the candidate's communicative ability and use of communication strategies. Being a monologue, it does not however measure interaction skills, questions and requests for repetition. By having either a peer candidate or the examiner asking following-up questions we may overcome this disadvantage, though.

The choice of topic is crucial and may favour one candidate over another. Clearly, in a talk on a given topic, it is important to have some factual content as well as opinions and reasonable arguments for that opinion. And even if raters are told not to stress the content of the speech, it is hardly possible to fully separate the form from its content. The preparation time gives rise to another problem: there is always the possibility that the candidates have learnt their presentations by heart. This pitfall may be avoided by shortening the preparation time. The most problematic issue with this test method, however, has to do with the kind of speech it elicits. As stated in Chapter 3, native speakers vary in their ability to structure long turns of factual information: even a perfectly fluent native speaker may find oral presentation of a topic difficult. This has to do with educational background, personality types and practice:

> The ability to construct […] long transactional turns appears to vary with individuals, in part, no doubt, depending on the opportunity they have had to produce long turns, which other people bother to listen to. The ability to produce long transactional turns, in which clear information is transferred is, we claim, not an ability which is automatically acquired by all native speakers of a language (Brown & Yule 1983:19)

It is in many cases ethically questionable to demand the students to perform a task that even native speakers would find hard to perform, as discussed earlier.

5.2.1.4 Role-play

Another common test method for oral production is the role-play: the test taker is asked to take on a particular role, and imagine himself in a given situation. Often he is asked to express different states of mind: anger, disappointment, content, joy etc. This technique is capable of eliciting special kinds of language use, which may otherwise be hard to elicit in a test. It is particularly well suited for eliciting language functions and fixed formulas, such as apologies, invitations, requests and denials, agreement and disagreement etc.

However, candidates have varying abilities to imagine their being someone else, at least in a test situation. Some people are reluctant to take on a role, while others find it amusing. In addition to personality differences, cultures differ in the degree to which they use role-plays in educational settings. Consequently, some candidates may be more familiar with the method than others. Moreover, some candidates may find it odd to express certain feelings towards the authority examiner, even in a role-play. This method may therefore work better between candidates than between a candidate and the examiner. In any case, when using the role-play, we run the risk of being measuring other skills than, or at least in addition to, the language proficiency of the candidates.

5.2.1.5 Reading aloud

In this technique the test taker is given some time to read through a text which he is then asked to read aloud for the examiner or rater. The obvious advantage of this method is that it is highly standardised and therefore easy to score. All test-takers get to perform exactly the same task, which makes it easy to compare performances across candidates. This should enhance the reliability of scores. In addition it is easy to administer and the test takers normally understand what is expected of them.

Reading aloud is not a communicative test method, however. It assesses only a limited part of the oral skill, namely pronunciation, but it does not yield information about the candidates' grammatical ability, the size of their vocabulary, not to mention their interactional skills. The most substantial criticism of this method, though, concerns the fact that it measures a skill irrelevant of oral proficiency, that is the ability to read aloud. Even native speakers vary considerably in their control of this skill, even though their pronunciation is without blemish. This again may put validity at stake.

5.2.1.6 Picture description

In this method the candidate is asked to describe a picture or tell a story related to a sequence of pictures. It is often used in relation to an interview or a discussion between candidates. When the examiner starts to ask questions about the picture, this tends to lead into the pattern of an interview and the candidate is less active (Underhill 1987:66).

As with the read-aloud test presented above, the picture description presents different candidates with the same prompt, and therefore the language elicited is more predictable and hence easier to judge than true free production. It is also suited for eliciting a particular type of vocabulary. Another advantage is that it can be used as a trigger for all candidates, also those who

have limited reading abilities, the illiterate and children, for whom a written text would be unsuited as a prompt for a following discussion or interview.

The drawback of this method is that people from different cultures tend to interpret the same pictures or stories in different ways. If a candidate misses the point, this may be a de-motivating start of an oral exam. The examiner or rater will interpret the picture within their cultural and personal frame of interpretation, and a candidate should not be penalised for interpreting the picture differently. The advantages of this method depend largely on the suitability of the picture or picture sequence used.

As shown in this chapter, the testing of oral ability has held fluctuating status over the last century due to difficulties in ensuring its reliability. Oral tests were common between 1600 and 1800 when it was replaced by written exams, all kinds of subjectively scored tests were avoided at the beginning of the psychometric-structuralist period with its focus on objectivity and reliability. Yet, later in the period, an attempt was made to reduce the subjective element of oral testing through a standardisation of the scoring procedure. A lot of work was put into the development of rating scales for oral proficiency, and training of raters in using these was highly recommended. In the communicative approach of today, the direct test of oral performance is considered the kind of test which best meets the demands of communicative testing. Yet there are challenges in relation to its scoring, which is the focus of the next two chapters.

Different test-methods elicit different kinds of speech. A read-aloud test, a role-play and a peer discussion give the candidates the opportunity to show different sides of their language skill. In addition, some techniques have the disadvantage of eliciting not only language proficiency but other skills or abilities as well. This is the case for read-aloud tests, role-play and oral presentation and may affect the construct validity of test scores. Test constructors need to be aware of the strengths and weaknesses of the test-methods they use when interpreting the scores of the test.

CHAPTER 6: RATING SCALES FOR ORAL PERFORMANCE TESTING.

To make sure that all test candidates are scored in the same way as far as possible, professional test constructors develop a plan for scoring common to all raters. This plan specifies rating criteria, level descriptions and scoring methods as well as the weighting and reporting of scores. This common practice of rating is what is referred to as the *rating procedure*. The distinct traits or components of performance upon which raters base their judgements are called *rating criteria*. Some typical criteria for oral performance are fluency, accuracy, organisation, and sociocultural appropriateness. (McNamara 2000:36). As the criteria are operationalisations of the underlying theoretical construct of the test, it is paramount that they be a true representation of the construct. If the criteria do not match the construct of the test, the test cannot be said to yield construct valid scores. Based on the rating criteria, raters place candidates on a common *scale*. Scales vary in the number of *levels or bands* they have: some scales have as few as two levels (*pass-fail*), while others have as many as 10 (American Council on the Teaching of Foreign Languages-(ACTFL) 1986). In addition, professional test constructors normally describe typical performance of the distinct levels of the scale, which serve as a guide for raters as well as feedback to test users. Such verbal descriptions of learner performance related to grades on a scale are called *level descriptors* or *band descriptors***.** An ordered series of level descriptors is called a *rating scale*.

This chapter addresses the issue of rating scales and rating criteria that raters use when assessing oral performance. It begins with a presentation of different kinds of rating scales, distinguished first by function and then by scoring method. Thereafter different sources of rating scale development are discussed: tradition, theory and empirical data respectively. Finally, some widely used criteria for oral assessment are briefly outlined.

6.1.Different kinds of rating scales.

There are different kinds of rating scales, and they may be classified according to the functions they fulfil, or according to the underlying scoring method of the test. In this section we shall start by looking at different functions of rating scales: as a raters' guide, as a test constructors' guide or as a report to test users. Thereafter we shall look at the differences between holistic and analytic rating scales in a linking of rating scale to scoring methods.

6.1.1 Different functions of rating scales

In the discussion of ratings and rating scales so far in this thesis, the main focus has been on rating scales as raters' guides in the process of judging language performance. However rating scales may fulfil other purposes as well. Alderson (1991a) identifies three different functions of rating scales: the function of guiding the test construction process, the function of guiding the rating process, and finally the function of reporting back to test users. Alderson argues that a rating scale devised with one particular purpose in mind may not always be used successfully for another purpose: For instance a rating scale devised as a raters' guide may be quite meaningless as a report to test users with no knowledge of linguistics or linguistic terminology. Test constructors should therefore be aware of the purpose they want their rating scale to fulfil before they construct the scale.

6.1.1.1 Guiding the test construction

The first, and maybe the least common, purpose of rating scales is that *of guiding the test construction*. Pollitt and Murray state that the purpose of such scales is "to describe the sorts of tasks that the student can do at each level, and so describe potential test items that might make up a discrete test for each level" (1996:74). Alderson mentions the ACTFL Guidelines as one example of a scale constructed with this intention. The proficiency of candidates at the distinct levels of the scale is seen in relation to the kinds of texts, tasks and items to which the candidates can respond successfully:

> Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning (Common Reference Levels: overall reading comprehension, C2, Common European Framework (CEF) of Reference for Languages: Learning, teaching, assessment 2001:69).

As the purpose of such scales is to guide the test constructors, Alderson labels this kind of scale constructor-oriented.

6.1.1.2 Guiding the raters

A more common use of rating scales is as *guidelines for raters*: When raters assess language production they always do so in accordance with a set of criteria. The criteria may be personal, internal and totally subjective, unconscious even, or as is more common in professional L2-testing, verbally explicit, common to all raters of the test and produced by the test constructors. The use of rating scales as a common standard for raters is a highly recommended and widely

used procedure in language testing and "a way of ensuring reliability as well as validity" (Alderson 1991a:73). These scales are called assessor-oriented, and are normally focused not so much on the different tasks and texts that candidates can perform, but rather on a candidate's relative control of different language components. The aim of assessor–oriented scales is to describe typical candidate performance at each level of the scale (Pollitt and Murray1996:74).

> […] Vocabulary is inadequate to express anything but elementary needs. Writing tends to be a loosely organised collection of sentence fragments on a very familiar topic. Makes continual errors in spelling, grammar and punctuation, but writing can be read and understood by a native speaker used to dealing with foreigners. Able to produce appropriately some fundamental sociolinguistic distinctions in formal and familiar style, such as appropriate subject pronouns, titles of address and basic social formulae (ACTFL provisional descriptors for writing, Intermediate-Low, referred to in Hughes 1989:90).

6.1.1.3 Reporting to test users

The third purpose of rating scales is to *report back to test users* (test-takers, teachers, admissions officers, employers etc.) the kinds of tasks, in work or social life, that a candidate at each level would be able to fulfil adequately (Pollitt and Murray 1996:74). The scale is a description of typical performance at different levels, which is easier for test-users to interpret than a single test score. This purpose of a scale is what can be termed user-oriented, and consequently the language used in the level descriptors must be less technical than in a constructor- or assessor-oriented scale. These scales are often holistic, offering one description at each level. An example of a holistic, user-oriented scale is presented below:

> Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/ herself and others and can ask and answer questions about personal details such as where he/ she lives, people he/ she knows and things he/ she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help (Common Reference Levels: global scale, A1, Common European Framework of Reference for Languages: Learning, teaching, assessment. 2001:24).

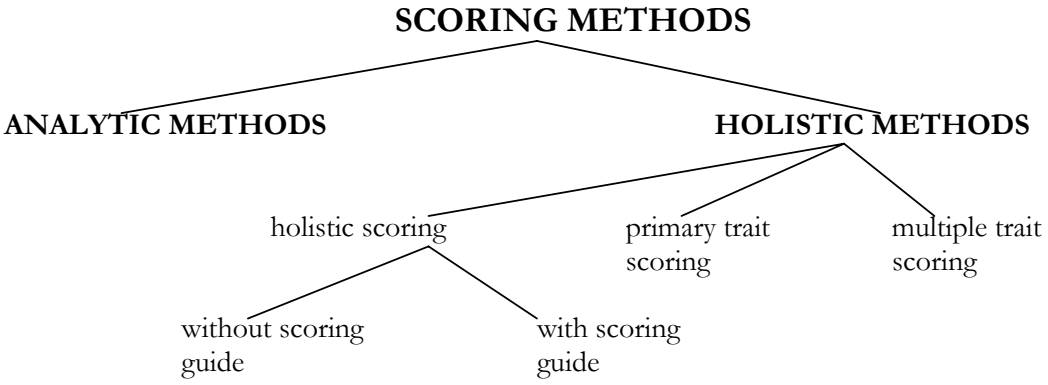
Ideally, in the construction, scoring and result-reporting of a single test, one should develop three kinds of scales to use in the distinct processes involved. Yet, in real life, it is more common that a scale developed with one purpose in mind is given other functions as well. For example assessor-oriented scales are sometimes given the reporting function, with the obvious problem of interpretability for test-users. It also occurs that scales are developed to fulfil the three functions simultaneously. This issue will not be pursued further here, though. In the following I shall focus on assessor-oriented scales exclusively, and on their effect on reliability and validity of oral test scores.

6.1.2 Holistic versus analytic or multiple trait scoring.

Different scoring methods also affect the shape and content of the rating scales. There are two main categories of scoring methods of performance tests: the holistic and the analytic method.

The *holistic scoring method* is traditionally referred to as the assignment of a single score to a sample of writing or speaking, based on the raters' impression of the performance as a whole rather than by focusing on its different sub-components. (Hughes 1989:86, ALTES' list of terms). The *analytic scoring method* on the other hand, implies the assignment of separate scores for a set of components of performance. (Hughes 1989, ALTE's list of terms). Both methods are used in the scoring of language production, writing or speaking. Hamp-Lyons (1991a) offers a more complex system of scoring categories, which is represented graphically in Figure 7:

Figure 7 System of scoring categories, Hamp-Lyons (1991a).



Hamp-Lyons (1991:243) maintains a clear distinction between analytic and holistic scoring methods based on whether or not the scoring requires the reader to "stop and count or tally incidents of the feature" (Cooper 1977:4). Holistic scoring methods are defined as all methods that do not involve the activity of counting incidents of wrong or correct language use. Hamp–Lyons' definition of holistic methods is therefore much wider than the traditional use of the term, and her definition of analytic scoring correspondingly narrow. In fact it is so narrow that very few scoring methods used today could be classified as analytic according to it. Hamp-Lyons argues for a sub-categorisation of holistic scoring methods into three different kinds: holistic, primary trait and multiple-trait scoring. What Hamp-Lyons calls *holistic scoring* is equivalent to the

definition of holistic scoring methods as the term is normally used and defined by Hughes, referred to above. Holistic scoring is either totally subjective when raters are not given any explicit rating criteria or rating scales to guide them, or focused when a rating guide is indeed available. In both cases each performance is assigned a single score.

Primary trait scoring is based on the view that the quality of the performance has to be seen in relation to the tasks candidates are asked to perform. It therefore implies that separate rating criteria be constructed for each task. This method is obviously very time consuming and is therefore mostly restricted to research contexts.

Multiple-trait scoring (MTS) is the last of the holistic scoring methods in Hamp-Lyons' overview and the one that she recommends for composition scoring in an ESL context. MTS requires raters to focus on several aspects of the performance and assign each trait a separate score, hence similar to the common use of the term analytic as defined by Hughes above. Hamp-Lyons, however, stresses that MTS is different from the analytic scoring methods "used in the 1960s and 1970s, which focused on relatively trivial features of texts (grammar, spelling, handwriting) and which did indeed reduce writing to an activity apparently composed of countable units strung together" (Hamp-Lyons 1991a:247). Yet, the term is currently used as equivalent to analytic scoring (Cumming 1997:54). When Hamp-Lyons chooses to maintain a distinction between MTS and analytic scoring[5], it is my view that this is due to theoretical rather than methodological differences between the two techniques. Both analytic and MTS methods require trained raters to focus on a set of traits of performance according to explicit rating criteria and to assign separate scores for each trait. The scoring method is consequently the same. Which traits raters are asked to emphasise, as well as the underlying view of language of the test constructors, are, as I see it, irrelevant to the scoring method. However, I find the term Multiple Trait Scoring favourable over the term analytic, firstly because of its transparency: it reveals quite clearly that the method involves focusing on various traits of performance. Secondly, because the term analytic is easily associated with the misleading idea that performance testing may be rendered objective. In this thesis, the term MTS is therefore used in Cummings's sense as synonymous with analytic and preferred to it, referring to all methods involving the assignment of several scores based on distinct traits of performance, irrespective of whether the traits are characterised as trivial or not. Holistic scoring methods are defined as the assignment of a single score based on raters' impressions of the text or speech sample as a whole.

[5] Personal e-mail communication 04.01

***6.1.2.1 Holistic and MTS rating scales*.**

Both MTS and holistic scoring methods may involve the use of explicit rating criteria and rating scales. The choice of scoring method will affect the characteristic of the rating scale.

A *holistic rating scale* is typically constructed by a set of global descriptors of performance at each level of the scale (Byrnes 1989). The descriptions may be restricted to a single expression as in the *English medium university test* referred to in Hughes (1989: 87):

> native speaker standard,
> close to native speaker standard
> clearly more than adequate
> possibly more than adequate
> adequate for study at this university
> doubtful
> clearly not adequate
> far below adequacy

Or it may consist of detailed descriptions of performance at each level. The ACTFL guidelines are one example of a holistic rating scale with detailed band descriptions (Hughes 1989, Saleva 1997). The following is an extract from the ACTFL guidelines referred to in Hughes (1989:90). It is a description of one of the six levels, which range from *Novice-Low* to *Intermediate-High*, passing through *Novice-Mid, Novice-High, Intermediate-Low* and *Intermediate-High*.

> Sufficient control of writing system to meet some survival needs and some limited social demands. Able to compose short paragraphs or take simple notes on very familiar topics grounded in personal experience. Can discuss likes and dislikes, daily routine, everyday events, and the like. Can express past time, using content words and time expressions, or with sporadically accurate verbs. Evidence of good control of basic constructions and inflections such as subject-verb agreement, noun-adjective agreement, and straightforward syntactic constructions in present or future time, though errors occasionally occur. May make frequent errors, however, when venturing beyond current level of linguistic competence. When resorting to a dictionary, often is unable to identify appropriate vocabulary, or uses dictionary entry in uninflected form (ACTFL Guidelines, description of level Intermediate-Mid, quoted in Hughes 1989:90)

The holistic rating scale takes as a starting point the scale levels, and describes the impression of performance often in relation to certain criteria of each level. In the extract of the ACTFL scale above, performance is evaluated according to what the candidate can do in writing, to grammatical control (past tenses, constructions, inflections) to the number of errors and to vocabulary. The holistic scale therefore implies that a candidate is at the same level in all the sub skills of language: In other words a Mid-Intermediate-candidate will be placed at this level in

grammatical control, vocabulary, the quantity of errors and in the ability to perform various tasks in writing.

A *MTS rating scale,* on the other hand, will start out with the criteria and describe a candidate's performance in relation to these. An MTS rating scale for a writing test may have the following criteria: grammar, vocabulary, mechanics, fluency and form. The candidate is evaluated on a scale from, say, 1-6 for each aspect according to a description of the levels related to each trait. The criterion *phonological control* is described as below in CEF (2001:117):

PHONOLOGICAL CONTROL	
C2	As C1
C1	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.
B2	Has acquired a clear, natural, pronunciation and intonation.
B1	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciation occur.
A2	Pronunciation is generally clear enough to be understood despite a noticeable foreign accent, but conversational partners will need to ask for repetition from time to time.
A1	Pronunciation of a very limited repertoire of learnt words and phrases can be understood with some effort by native speakers used to dealing with speakers of his/ her language group.

This implies that a candidate may get distinct scores in the distinct components of the skill measured. The final scores of an MTS scale may be reported separately or summed up and reported as a single score. Assessor-oriented scales may be holistic or MTS, but most scales with the purpose of guiding assessors are of the last kind (Pollitt and Murray 1996).

A final remark about holistic and MTS methods regards the advantages and disadvantages of the two methods. The main advantage of the holistic scoring method is that it is quick and therefore economic. It often takes experienced raters only a couple of minutes to gain a general impression of a composition, while the detailed analysis of sub- components involved in MTS scoring is a time-consuming activity. Proponents of the holistic scoring method argue that methods focusing on sub-components reduce the raters' full perception of the performance (White 1985), while Hamp-Lyons opposes this view, arguing that it is holistic scoring which is reductive, "reducing the writer's cognitively and linguistically complex responses to a single score" (Hamp-Lyons

1991a: 244). Another serious drawback connected to holistic scoring, especially apparent in a SL context, is the fact that learner language develops in uneven paths: L2 learners often acquire different sub-skills at different rates. The nature of learner language development is more truly echoed through the use of MTS rating scales.

An additional advantage of MTS is that the scores may fulfil a diagnostic purpose. Candidates may get to know which aspects of language they need to work more on in quite a different way than is made possible through holistic scoring.

Today there is consensus in the field that language ability is a many-faceted skill consisting of several sub components (Bachman 1990). In such a view of language, it is natural that the scoring method and rating scale reflect these sub-components.

Finally, the use of MTS and specifications of sub-skills, should enhance both the reliability and the construct validity of test scores (Alderson 1991a). The focus on distinct sub-components specified by test constructors should affect rater-reliability positively: it seems like a reasonable assumption that raters score more in agreement with each other when focusing on the same set of explicit criteria than when relying on their individual gut feelings. As the rating criteria are operationalisations of the test construct, the use of explicit rating criteria should ensure that raters focus on the components that test constructors consider relevant to language performance and not on irrelevant constructs such as cultural or national background, sex, personality etc. This hypothesis needs to be investigated empirically and indeed, it is one of the main hypotheses of the present project.

6.2 Approaches to rating scale development

So far, we have seen that rating scales may fulfil different purposes, and they may represent different scoring methods. Another relevant question in relation to rating scales is: how are they constructed? The sources of the rating scales are often unknown, which leaves us with the limited option of hypothesising as to the base upon which the scale is built (Brindley 1998:117). Furthermore there are often multiple sources of a rating scale: when constructing rating scales, test constructors may use their own intuition as native speakers, their experience with learner language and previous scale-construction as well as linguistic models and empirical data in combination. Probably, scales are also based on well-established rating scales, of which the sources may be uncertain. The overview of sources of rating scales presented in this section, is therefore necessarily an oversimplification. Despite the lack of real clear-cut categories, I have chosen to group rating scales according to three sources apparent in the LT literature: the FSI-

tradition, linguistic theory and empirical data. The last two approaches may be further subcategorised according to the kinds of theory and empirical data upon which they are built.

6.2.1 The FSI/ ILR/ ACTFL- or the traditional approach

The FSI/ ILR/ ACTFL- or *the traditional approach* is one example of a scale development whose source is uncertain. References to its base in the LT literature are inconsistent. Fulcher argues that in the traditional approach:

> […] the descriptors of the rating scales are constructed by an expert, often using his or her intuitive judgement concerning the nature of developing language proficiency […] [Such scales] mostly have in common the lack of any empirical underpinning […] (Fulcher 1996a: 208).

However, Liskin–Gasparro (1984: 37, referred to in Brindley 1998: 117) argues for the opposite view, i.e. that the ACTFL guidelines

> […] were developed empirically, that is by observing how second language learners progress in the functions they can express or comprehend, the topic areas they can deal with, and the accuracy with which they receive or convey a message. The guidelines then, are descriptive rather than prescriptive, based on experience rather than theory.

One reason why the source of the FSI/ ILR/ ACTFL scale is blurry may be the simple fact that it was constructed in what Spolsky refers to as the "pre-scientific" period, that is before language testing emerged as a professional field. (Spolsky 1995). It had its genesis in the United States in the 1950s in connection with the Second World War and the growing concern about the language proficiency of US diplomats and civil servants in overseas postings. As outlined in the previous chapter, the Foreign Service Institute (FSI) was bestowed with the task of developing an oral interview test and related rating scale. The rating scale was developed by a committee in consultation with John B. Carroll. As the FSI interview and rating scale spread to other government agencies, the level descriptors used by different institutions were standardised under the name of The Interagency Language Roundtable (ILR) Oral Proficiency Scale in 1968. In the 1970s the use of the ILR interview spread outside the federal agencies, and ACTFL was put in charge of its further development. New guidelines for the test were developed in 1982 and 86, and theoretical innovations such as the development of a theoretical foundation of communicative competence in the 1980 were to some extent incorporated (Spolsky 1995:178). The FSI/ ILR/ ACTFL rating scale has been very influential in scale development in the US as well as outside (Alderson 1991a:71).

6.2.2 The theory-driven approach

The *theory-driven* approach refers to scale developments that use linguistic theory as their point of departure. The theoretical base may be general models of language ability such as Bachman's model of communicative language ability (CLA) (Bachman 1990), or it may be theories which describe characteristics of spoken interaction in particular, such as, for example, Bygate's model of speech or other discourse models. In addition it may be theories of language development of non-native speakers, i.e. theories of second language acquisition, such as for example the ZISA group's Multidimensional Model (Meisel et al 1981).

6.2.2.1 Based on general models of language ability

Theory-driven approaches to scale development have traditionally been based upon *general models of language ability*. Bachman's model of communicative language ability (CLA) and the subsequent Bachman and Palmer (1996) model have been particularly influential maybe because it was developed by language test theoreticians for the purpose of test construction. An obvious problem with these models, however, is that they do not distinguish characteristics of the skill of speaking as opposed to the skill of writing. This has lead test constructors to develop their own models of spoken language ability building on these models (Milanovic et al. 1995, Saville and Hargreaves 1999). The models of CC may be helpful in selecting which criteria to include in the scales, but their value is limited when it comes to defining and describing the bands of the rating scales.

6.2.2.2 Based on models of spoken interaction.

Saleva (1997) stresses the importance of using rating criteria and scale descriptors based on *models of spoken interaction*. She argues that the comparison of a test takers' oral performance to planned or written texts is unreasonable, as oral production is restricted by conditions different to those confining written production and seems less perfect than written language (see Chapter three). Saleva argues that: "learner talk in natural conversation […] should not be compared to standard English based on planned or written text but with natural native speaker conversations" (Saleva 1997:23). Fulcher too emphasises the promising role of discourse analytic models as sources of scale development (Fulcher 1997), as does Hasselgren (1998). The use of discourse analysis in oral scale construction is not well established in the field despite its obvious advantages over general models of language ability.

6.2.2.3 Based on SLA-theories

A final approach which has been very little investigated in the field, is the use of *SLA- theories* as a basis for scale development. Traditionally the contact between the fields of LT and that of SLA has been of limited scope, but this is now beginning to change (Bachman and Cohen 1998:1). De Jong claims that:

> What we need to know if we want to develop good scales is not linguistic knowledge of how language is structured, what all the features of language are; we need to know how somebody acquires language, that is, what the developmental stages in language acquisition are (de Jong 1988: 74).

The call for validation of rating scales against theories of language acquisition is also made by Kaftandjieva and Takala 2003:31.

The Multidimensional Model (MM) developed by the ZISA-group in the 1980s (Meisel et al 1981) and further developed in Processability Theory (PT) (Pienemann 1998) is one possible SLA-source for scale development. These theories make a claim that a number of grammatical structures, such as word order and some grammatical morphemes are acquired by learners in developmental sequences. They could therefore be used in the construction of an MTS rating scale for the grammatical component of speech. If the SLA approach should be fully appreciated, however, similar models of developmental stages would have to be developed for the other components of speech such as fluency, vocabulary, pronunciation, comprehension etc. Hasselgren's work on small-words and fluency is an important contribution in this respect (Hasselgren 1998). Hasselgren's work is classified under the section of data-driven approaches to scale-development below and will be further commented on there. It is mentioned here because is meets de Jong's demands for building rating scales from SLA insights.

6.2.3. The data-driven approach

In contrast to scale development approaches built on theory of language, we find approaches that use empirical data as the point of departure. The chief proponents of this *data driven approach* is Fulcher (1993, 1996, 1997) who criticises the traditional theory-driven approach to scale development for being "[…] a circular, self-contained notion of scale development which appears to lack empirical support or theoretical credibility" (1996:212). The data-driven approach may be further subcategorised according to the kind of data used as its basis: non-native speaker performance, native speaker performance or raters' evaluations of candidate performance.

6.2.3.1 Based on non-native speakers' performance

Fulcher has been particularly interested in developing rating scales for fluency. In respect to this he argues that "the definitions of fluency which exist seem to be inadequate for the purpose of operationalisation in a test […]" (Fulcher 1996a: 210), and the solution he offers is to base the band descriptors upon an analysis of *non-native speakers' (NNS) performance* (Fulcher 1993, 1996). By using this approach, Fulcher claims to avoid two common pitfalls in scale development: i.e. the lack of detailed definitions of abilities makes it impossible to investigate the validity of the scale, and besides, the descriptors are not specific enough to relate to actual performance (Fulcher 1996a:225). By using NNS' performance data as a starting point for scale development, Fulcher develops detailed descriptors of performance at different levels, which not only serves the purpose of raters' guides but may also be the subject of validation studies.

6.2.3.2 Based on a comparison of NNS' and NS' performance

Like Fulcher, Hasselgren (1998) also uses empirical data as a starting point for band descriptors of fluency. In addition to the use of non-native speakers as informants, Hasselgren uses a group of *native speakers (NS)* in her study. Hasselgren investigates the use and frequency of small words in speech produced by two NNS groups at varying performance levels and a NS group. Her findings are then used for developing descriptors of performance at three levels (Hasselgren 1998:274-75).

6.2.3.3 Based on raters' perception of NNS' performance

Other scholars arguing for an empirical base for scale development are Pollitt and Murray (1996). In contrast to Fulcher and Hasselgren, Pollitt and Murray claim that scale descriptors should be based, not on candidates' performance, but on *raters' perception of this performance*. They stress two advantages of this approach: the rating scales will be more accurate if they focus on aspects salient in performance at each level of performance, and they will be easier to apply if they are based on the raters' own perceptions (Pollitt and Murray 1996:89). This method has some serious drawbacks however: Research has shown that different rater groups emphasise distinct aspects of candidate performance (Brown 1995, Brindley 1991, Chalhoub Deville 1995). Pollitt and Murray's scales would therefore probably be qualitatively different depending on whether raters were naïve native speakers, S2-teachers, experienced linguists or trained raters. Another serious problem is that this method implies a confidence in raters' judgements which may very well be unjustified: as Pollitt and Murray recognise, raters in their study were sometimes positively wrong in their judgements (Pollitt and Murray 1996:87): Raters noted aspects of performance that were

not actually present in the speech sample they evaluated. Furthermore, they were influenced by features irrelevant to language ability such as "the candidates' personalities, physical attractiveness, nationalities and cultural background" (Pollitt and Murray 1996:87). When Pollitt and Murray reject these aspects as irrelevant to the "official construct" of the test, the method is reduced to an analysis of the match between the official construct (present in the researchers' mind) and the raters' internal constructs. Whenever there is a mismatch between the two constructs, it seems that the official construct is given priority. It is therefore questionable whether this approach is really data based at all, or whether data are used merely for underpinning the researchers' hypotheses of the construct of oral proficiency.

It is obviously hard to decide which of these approaches are the most fruitful for scale construction. It will probably depend upon the purpose of the test. In my view, a scale based on a solid theory of spoken language would be the most powerful approach. If the theory has been tested against empirical data, the scale would be as valid as if it was based on data alone. Whether it should be based on general models of language ability, interactional models or models of SLA would also have to be considered in relation to the purpose of the test. In some academic language tests, the interactional aspect is less pronounced than in a test at lower or intermediate levels. In the academic test, then, a general model could probably be used as a basis, while the lower to intermediate level would profit from reference to theories of spoken interaction. SLA models clearly have potential as a basis for scale construction. We would however need similar descriptions of acquisition order for the other aspects of language as well for this approach to be fully appreciated. Whatever approach for scale construction, the validity of the scale has to be established before the scale is used in live tests (Kaftandjieva and Takala 2003). Since the validity of the NORS is taken as a premise in this thesis, I will not go in detail on this here.

Summing up, then, we have seen that a rating scale may fulfil different purposes. It may be used as guidance in the test construction process, it may be used as a raters' guide, or it may be given the role of reporting back the interpretation of the grades to test users. A scale constructed for one purpose may not always be suitable for another. It is therefore crucial that test constructors be conscious of the purpose of their scales in scale construction, though this seems not always to be the case.

Thereafter rating scales were discussed in relation to two kinds of scoring methods, the holistic and the analytic or multiple trait methods. It was demonstrated how the choice of scoring

method affects the scales, and advantages and disadvantages of holistic versus multiple trait rating scales were discussed.

Finally, we looked at the different sources for scale development. There has been a multitude of different approaches to scale development, and there is a lack of consensus in the field as to whether rating scales for NNS' oral performance should be based on other well established scales, on theory or on empirical data.

CHAPTER 7: THE RATER

As repeated throughout this thesis, tests of oral performance cannot be scored satisfactorily without the use of human raters. The rater nevertheless represents a potential source of measurement error (Bachman 1990, Bachman et al 1995, McNamara 1996). As discussed in Chapter 5, the difficulties in ensuring the reliability of test scores based on raters' judgements, were in some periods of the history of language testing considered insurmountable and lead to the exclusion of oral assessment from language tests altogether. In the communicative paradigm of today, professional language testers agree that the ability to speak is a central language skill, which should be measured despite the potential unreliability of the subjective judgement involved.

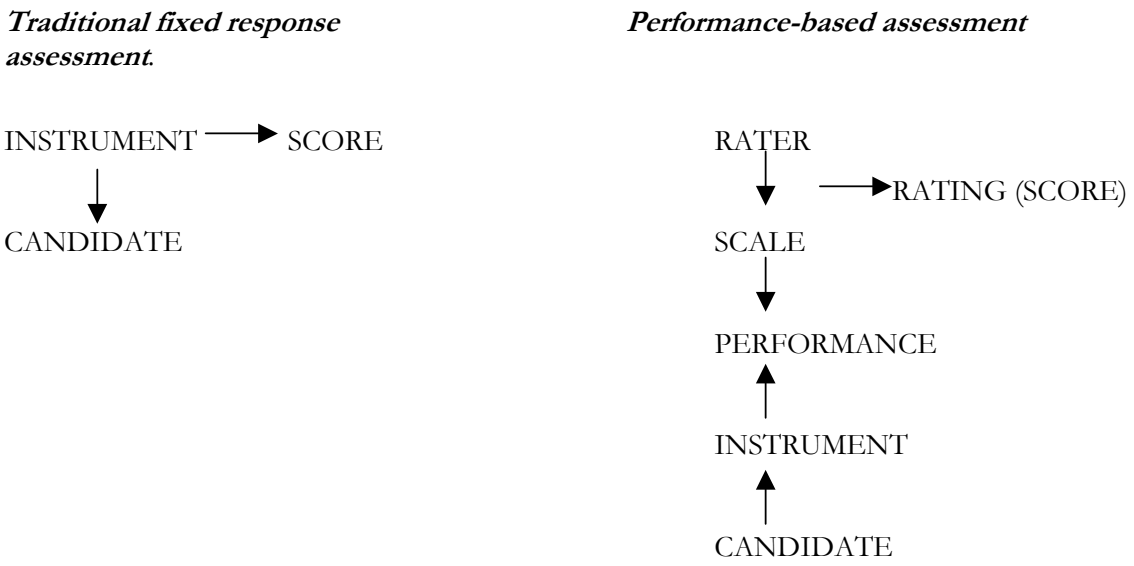
The consistency of rater-based scores can be heightened by the use of some highly recommended and frequently used procedures: firstly, test constructors should develop a set of criteria and an explicit rating scale for raters to use as a basis for their evaluation. (Rating scales for oral tests were discussed in detail in Chapter 6 and will not be further covered again here). Secondly, raters should receive training in how to interpret and use the scale. And finally, since one cannot completely eliminate the discrepancy between raters, each candidate should be scored by at least two independent raters whose scores should be averaged (Alderson 1991b).

The subject of the present chapter is the rater variable and its effect on reliability as well as validity of test scores. I start by introducing two models placing the rater variable within a theoretical framework. Building on these, I present a third model, which I use as a basis for the further treatment of the rater. Thereafter follows a discussion of the effect of the rater variable on reliability and validity before the purpose of rater training is discussed. Finally, I give an overview of recent rater-related research.

7.1 Theoretical framework for the rater variable

In his 1996 book McNamara introduces a model of performance tests based on Kenyon (1992, referred to in McNamara 1996), which places the rater variable within a theoretical framework.

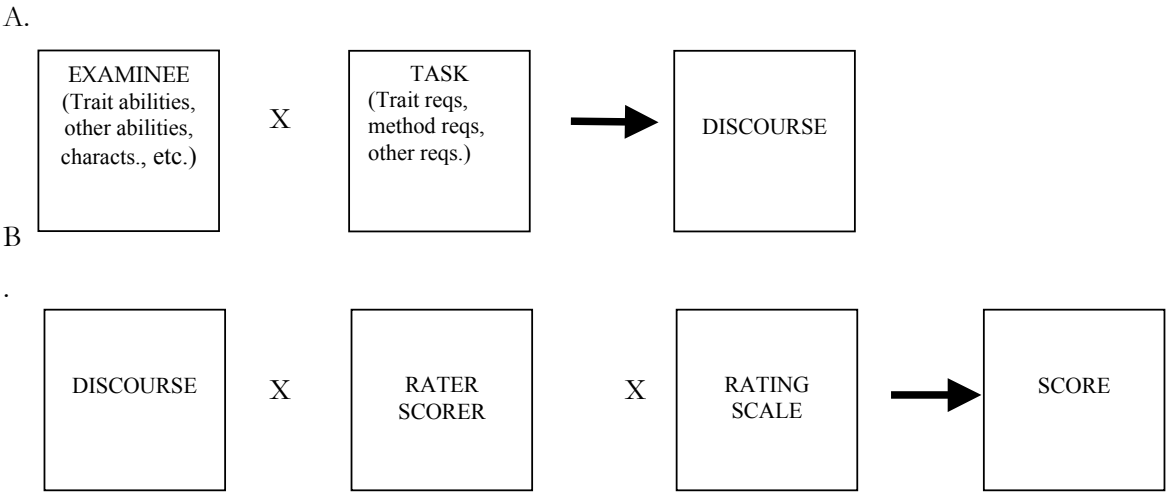
Figure 8 Characteristics of performance based assessment, McNamara 1996.



McNamara argues that "The format of a performance-based assessment is distinguished from the traditional assessment by the presence of two factors: a *performance* by the candidate which is observed and judged using an agreed *judging process*" (McNamara 1996:10). One may however correctly maintain that the agreed scoring procedure is an ideal but not a necessary condition for performance tests. Indeed, there are numerous examples of performance tests in Norwegian society where a candidate's performance is judged by one or several raters, but where an agreed scoring procedure and common rating scale are lacking. (see Section5.1.1). I would therefore argue that the presence of raters together with the performance produced by the candidates are the two principal characteristics of performance tests, while an agreed scoring procedure is an ideal, but not a condition, for performance tests.

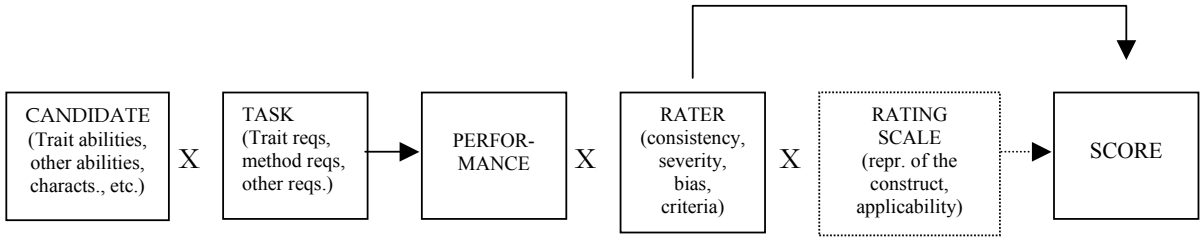
A similar model is developed by Upshur and Turner (1999).

Figure 9 Characteristics of performance test taking and scoring, Upshur and Turner, 1999.



The main difference between McNamara's model and Upshur and Turner's model is that the latter includes the abilities and requirements necessary for the candidates in order to perform the task. The model is based on the thinking of the True Score Model, which maintains that "test performance depends […] upon a number of non-trait abilities of the test taker. These constitute systematic errors of measurement" (Upshur and Turner 1999:84). Upshur and Turner acknowledge that the rater variable constitutes a potential source of measurement error (Upshur and Turner 1999:85). This knowledge is not, however, fully represented in their model: the boxes containing the rater and the rating scale lack reference to abilities and requirements of raters and scales necessary for the test to yield reliable and valid scores. I have included this factor in an elaborated model of performance based tests, presented in Figure 10 below.

Figure 10 Extended model of performance based tests.



Like the Upshur & Turner and McNamrara models, this model takes account of three characteristics of performance tests: *performance, rater* and *rating scale*. It differs from the previous models by taking into consideration the lack of an agreed scoring method, which characterises many performance-based tests around the world. It is an extension of the earlier models by

including rater and rating scale characteristics necessary for the test to yield reliable and valid scores. For a *rating scale* to be valid, it has to represent the construct in a satisfactory way, it should be clear in its specifications and easy to interpret and apply by the raters, as discussed in detail in Chapter 6. *Raters* should be internally consistent, they should to some degree rate in accordance with each other, their scoring should not be in favour of some candidate groups at the expense of others, that is they should not be biased. Finally, raters should base their scores on the criteria explicit in the rating scale and not on other individual criteria. We shall return to this point in the treatment of the rater effect on reliability and validity in the next section.

7.2 The rater effect on test scores

It has long been recognised that the rater variable affects the scores of a test. Linacre (1989) (referred to in McNamara 1996:122) argues that the rater effect accounts for as much variation in test scores as variation due to the candidates' degrees of proficiency. In traditional language testing, the rater variable has above all been related to the reliability or consistency of test scores. Performance based tests have been taken to hold high validity, almost by nature (Huot 1990: 204). However, in modern test research there is a growing awareness of the fact that the rater variable also affects the construct validity of test scores. We shall approach these two issues in turn.

7.2.1 The rater effect on reliability

As discussed in Section 4.2.1, the test literature distinguishes between two kinds of rater related reliability: intra-rater and inter-rater reliability. *Intra-rater reliability* is defined as an "[…] estimate of the reliability of assessment, based on the degree to which the same assessor scores the same performances similarly on different occasions" (ALTE's list of terms 1998). It is an estimation of the internal consistency of each and the same rater, and investigates questions such as whether different candidates are scored consistently by the same rater.

Inter-rater reliability (IRR), on the other hand, is defined as an "[…] estimate of test reliability based on the degree to which different assessors agree in their assessment of candidates' performance" (ALTE's list of terms 1998). Inter-rater reliability, then, has to do with homogeneity within the group of raters, and a relevant question is whether the same candidate is assessed alike by different raters.

In Chapter 4 classical test theory versus item response theory, (with the related true score model versus multi-faceted Rasch analysis) were discussed. These two approaches entail qualitatively

different approaches to the study of the rater variable. In classical test theory it is an explicit aim to reduce error score: rater related variance due to internal inconsistencies of each rater, as well as disagreement between raters, are sources of measurement error and should therefore be reduced. Reliability coefficients close to 1.00 are considered the ideal as it means that the error score due to the rater variable has been eliminated.

Within the framework of *ITR,* on the other hand, inter-rater agreement is no longer considered worthy of research concern. On the contrary, a certain degree of variation between raters is seen as necessary in assuring test validity: communicative language ability in speaking and writing is multi-componential and so intricate that a complete agreement between raters would jeopardise a true evaluation of all aspects of the skill, it is argued. Huot (1990:210) claims that "[…] making raters judge alike could reduce their powers of perception". Rather than forcing raters to agree, one should focus on heightening intra-rater reliability, a point also made by McNamara (1996:232): "[…] the traditional aim of rater training- to eliminate as far as possible differences between raters- is unachievable and possibly undesirable". Yet, he continues " […] we should not conclude that rater training is a waste of time; […] it helps raters achieve *self*-consistency, the *sine qua non* of acceptable rater behaviour" (ibid). McNamara's claim is echoed by Alderson:

> […] I find it easier to conceptualise a rater's agreement with his or her judgement on some precious occasion as indicating reliability than I do the notion of one rater agreeing with another, who brings different intellectual and experiential baggage to the rating process". (Alderson 1991b:64).

A second reason why the importance of rater-agreement is diminished in IRT is due to statistical innovation in this approach. Multi-faceted Rasch analysis and the data program FACETS (Linacre 1989) may actually take differences between raters into account and correct for such variability. Consequently, raters need not be equally severe, as long as they are internally consistent.

7.2.2. The rater effect on validity

The rater effect on test scores has traditionally been associated with a potential lack of test reliability, and the challenge when using subjective evaluation has for that reason been to enhance reliability (Spolsky 1995:99). Yet, as Spolsky claims, "[…] reliability proved to be a jealous god, making it hard to pay attention to other demands" (Spolsky 1995:193). One of these "other demands" was the aspect of validity. Huot argues along these lines in his call for focus on validity aspects in holistic scoring (the same should hold true for MTS as well):

> While the emphasis on reliability was a necessary phase in the growth of holistic scoring, I believe it is time for the profession to begin to ask questions about the validity of holistic scoring methods. The unquestioned assumptions, confusion, and neglect of validity in holistic scoring have gone on long enough (Huot 1990:211).

Huot's appeal has been widely recognised during the last decade. In modern language test research there is a growing understanding that the effect of the rater variable is not restricted to reliability, but concerns the very validity of scores. Weigle argues that:

> In a criterion-referenced test, where a given score is associated with a verbal description of observable behavior, the extent to which raters can agree on this description is critical to the validity of inferences based on test scores (Weigle 1994: 203).

This makes sense: as stated in Chapter 6, for a subjectively scored test to yield valid scores, the rating scale has to be a valid operationalisation of the underlying construct of the test. If, for example, a test is assumed to elicit organisational as well as pragmatic competencies, the rating scale should reflect both aspects. If it focuses exclusively on formal linguistic correctness, it cannot be said to yield valid scores. But, surely, the case would be no different if the scale correctly focuses on both aspects, but raters nevertheless fail to pay attention to pragmatic aspects. The match between the internal scores of the raters and the criteria of the rating scale is therefore of crucial importance to construct validity of performance based tests.

In recent test research and literature, the effect of the rater variable on construct validity has been repeatedly stressed (Vaughan 1991, Connor-Linton 1995, Milanovic et al 1996, Cushing Weigle 1994, 1998, Shi 2001, Alderson and Banajeree 2001), and the value of studies focusing on the effect of the rater variable on reliability in isolation, has been questioned. Such studies are insufficient, as argued by Shohamy et al in 1992:

> […] the degree of agreement obtained by the various raters, the degree of stability, or any other type of reliability, does not provide any information about the validity of the test. Raters of different backgrounds and training may agree on the "wrong" things. Thus, while reliability is a necessary condition of writing tests, it is insufficient; research on the validity of the writing tests should be conducted as well (Shohamy et al 1992:32).

This claim is echoed by Connor-Linton who argues that "quantitative similarities in ratings may mask significant qualitative differences in the reasons for those ratings" (Connor-Linton 1995: 99). In other words, high reliability coefficients may mask the fact that raters may have arrived at those similar scores by different routes, they may focus on different aspects of performance, which would jeopardise the construct validity of test scores. We shall look more closely at some of the studies of the rater effect on test scores in 7.4.

7.2.3 Inter-rater reliability as validity, or why not?

In Sections 7.2.1 and 7.2.2 the rater effect on reliability and then on validity was dealt with in turn. This division is based on a conventional assumption in language testing that reliability and validity are distinct features of test scores: when investigating reliability, we gather information about the stability or consistency of tests scores. When studying validity, on the other hand, we focus on the evidences produced by the test and the certainty with which we may base our inferences on test results. Traditionally, language test theory has drawn a sharp line between reliability and validity: when investigating reliability we are interested in gathering information about the stability or consistency of tests scores, when studying validity we focus on the evidences produced by the test and the certainty with which we may base our inferences on test results. The aim of reliability studies is to describe and control the factors producing error score, while the purpose of validity studies is to describe what is left when the error score is reduced (Bachman 1990). Reliability is seen as a necessary but not sufficient condition for validity, a test may produce reliable but not valid scores, while the opposite is not the case (Johnsen 2003).

While it is rather unproblematic to maintain a clear distinction between the concepts for objectively scored tests, the distinction becomes blurred for subjectively scored tests (Bachman 1990). Within the unified concept of construct validity presented above, it may even be reasonable to argue for rater related reliability as a kind of validity. Alderson (1991b) argues that when raters disagree, it is impossible to deduce whether this is a problem of reliability or validity:

> And if raters produce different ratings, is this because they are unreliable, or because they are reliably using the wrong answer key- that is, they are applying the criteria differently (are misinterpreting the criteria, are using their own internalised interpretations of the words and descriptions, etc). […] One may thus again be led to see the inter-rater reliability coefficient as an index of validity, not reliability (Alderson 1991b:64).

Disagreement between raters about the scores may be due to two factors: they may agree about the construct and base their scores on the same set of criteria, yet still be differently severe. Or, on the other hand, they may be equally severe but focus on different aspects of the performance. When faced with rater disagreement it is not evident from the scores alone which of the two factors caused the disagreement. However, the sources of disagreement are indeed distinct, and they may be revealed through research, of which this project is only one example. I will therefore maintain a distinction between inter-rater reliability and validity, defining inter-rater reliability as agreement between raters about the grade scores, and validity, in this context, as the match between the criteria used by raters and those explicit in the rating scale of the test in question.

In a later article, Alderson calls his earlier concern as to whether inter-rater agreement should be classified as reliability or validity, an unnecessary agony. Within the frame of a unified concept of validity as presented in Chapter 4, the importance, he argues, lies in identifying and reducing variability in test scores. How we chose to label these different sources of variability is less important. In fact, Alderson argues that "[…] making a distinction between "reliability" and "validity" is irrelevant in this unified view of validity" (Alderson and Banerjee 2001:102).

I would nevertheless argue that there *are* different sources of rater disagreement, and even though they are not immediately observable on the basis of test scores alone, they may be revealed through research (which this project is only one of several examples of). I will therefore maintain a distinction between inter-rater reliability and validity in the present work. Inter-rater reliability is defined as the mere agreement between raters about the grade scores without reference to the underlying criteria for these scores, while construct validity is defined as the match between raters' criteria and those of the rating scale

7.3 Rater training

One of the main purposes of the present project has been to investigate the effect of rater training on reliability and validity of test scores. But what is rater training, what is its purposes and how is it conducted? What do we know about its effects on test scores?

Rater training is the schooling of people, often language teachers, in how to judge oral or written language performances. It has been a recommended procedure for rater-based tests in the US since the 1950s (Spolsky 1995) and it is increasingly being used in language testing in Europe and elsewhere as language testing reaches a higher level of professionalisation. However, it is not yet a standard procedure in Norway. With a few exceptions, scores are mostly based on raters' subjective evaluations. Raters' experience, rather than high quality rating scales and rater training, is used as a justification of the quality of scorings.

7.3.1 The purpose of rater training.

The purpose of rater training is naturally closely related to the concept of rater effect on test scores: if the underlying assumption is that the rater variable affects test reliability, a purpose of the training session will be to make raters internally consistent as well as in agreement with each other. If, on the other hand, one assumes the rater variable to affect construct validity as well, a

chief goal will be to make raters focus on the aspects of performance explicit in the rating scales. The purposes of rater training are therefore diverse:

- to enhance inter-rater reliability
- to enhance intra-rater reliability
- to enhance construct validity
- to make raters able to defend their scores

The most widely accepted and frequently referred to purpose of the training sessions is to heighten *inter-rater reliability* (Weir 1990:80, McNamara 1996, Lumley and McNamara 1995:56). For a test to be reliable, raters should assign the same candidate with approximately the same score. That is, raters should be equally severe and they have to apply the grade scale in similar ways. It is not enough that they are able to describe performance according to a set of criteria; in addition they need to be able to connect these different ability levels to the same levels on a grade scale.

Within the framework of IRT and MFR, inter-rater reliability is, as discussed earlier, given little consideration. When rater training is still considered important, this is due to other factors than the demand of augmenting rater agreement. A second purpose of rater training, which is embraced in modern approaches to language testing as well, is to make each rater internally consistent, that is to heighten *intra-rater reliability*. As Cushing Weigle argues: "[…] the function of training is not, or should not necessarily be, to force raters into agreement with each other (inter-rater reliability), but rather to train raters to be self-consistent (intra-rater reliability)" (Cushing Weigle 1998:264). Lack of inter-rater reliability, due to the fact that some raters are consistently more harsh or lenient than others, may be corrected for by using MFR. It is not, however, possible to correct for the lack of rater internal consistency.

A function of rater training, which has gained increasing focus during the last decade, is to enhance *construct validity* of scores. Rater training does not only relate to numeric scores and rater severity. Just as important is its effect on the criteria raters use as a basis for the scores. This is a main research question of the present project, as it was in the work of Cushing Weigle who poses the important question: "[…] to what extent does rater training function to bring raters into agreement about the definition of the ability which the test is intended to measure?" (Cushing Weigle 1998:265).

A final purpose of rater training is to raise raters' consciousness about the scoring process. For many high-stake tests, candidates may complain and demand an explanation for their scores. Through careful and attentive work with the criteria and band descriptors, raters should be more capable of *defending their scores* and give relevant explanations to the candidates. A

thorough understanding of the underlying construct of the test, and the ability to verbalise this, is also an advantage for teachers' diagnosis of and feedback to their pupils. Rater training should therefore also have a positive side effect for teachers who participate in it.

7.3.2 The procedure of rater training

The *procedure of rater training* will vary in accordance with the purpose of rater training considered most important by the test constructors. Rater training typically involves raters getting together with test designers in group-sessions. Test designers will explain the specifications and scale descriptors and how the different parts should be interpreted. An important part of the session is consciousness-raising about the underlying construct of the test. Discussions about the criteria and rating scale descriptors are important in this respect. Raters are presented with a set of oral or written L2 language samples, and use the rating scale in assessing the candidates. Discussion of actual performances and how to apply the scale in assessment of these performances are important ingredients in rater training sessions. For writing, it is also possible to have raters score a set of performances prior to the training session. Test constructors may conduct different kinds of analysis based on these scores, and report the results to the raters on the session. It may be an aid for the raters to know if they are harsher or more lenient than the average rater of the group, for example.

Despite the fact that rater training has been a highly recommended procedure for quite some time, it has not, surprisingly enough, been subject to much research interest (McNamara 1996:126). Some studies have been conducted, though: the effect of rater training on test scores will be handled as part of the next section focusing on rater related research.

Before leaving rater training aside for a moment, it is necessary to focus on one important distinction in relation to trained raters. The test literature and rater related research do not always makes a distinction between trained and experienced raters. Rather the characteristics "trained and experienced" are grouped as one category (McNamara 1996, Weigle 1998). Weigle argues that the two should not be separated, as "the phrase *training process* refers both to the formal training received in the norming session, and to the informal training received through the live rating"(Weigle 1998:267). The training sessions are only one part of the rater training, the other part is the participation in live scorings and the following discussion between raters. This is the way the category term *trained raters* is used in this work, as well.

7.4 Previous research on the rater variable

The rater variable has received growing research interest during the last decade. There are several possible reasons for this. Firstly, there has been a growing interest for and use of subjectively scored tests during the last decade (Weir 1990, Spolsky 1995, McNamara 1996). Since the rater variable is a potential source of measurement error, the increase in use of performance tests naturally leads to an increased need for understanding and controlling the potential variability due to the rater effect. As McNamara states:

> In order to have confidence in the rating procedures for which we may be responsible, or whose results we may be relying on, it is necessary to investigate and control for the effect of factors such as [*raters, rating procedure, task choice and interlocutors*] (McNamara 1997:134).

A second cause may be the methodological and statistical innovations in item response theory such as the development of MFR. These approaches allow new and interesting research questions to be raised as to the interaction between the rater variable and other variables affecting the score on performance tests (McNamara 1996, Weigle 1994, Bachman, Lynch and Mason 1995, Lynch and McNamara 1998).

A third reason for the increased interest in the rater variable is the growing understanding that the rater variable affects the construct validity of the test, as discussed in Section 7.2.2. If raters use different criteria from the ones expressed in the rating scale the test will fail to measure what it intends to measure. In modern language testing, validity is considered a prime virtue of test scores and of paramount interest to test researchers. When the rater variable is assumed to affect validity and not only reliability as in earlier language test history, it becomes more interesting as a subject of research focus.

In this section I present an overview of rater-related research. The approaches, as well as the research questions, are multiple and diverse: some studies focus on the effect of the rater variable on reliability alone (Shohamy 1983, Shohamy et al 1992, Sieloff Magnan 1987), while the majority of rater-related studies during the last decade look into the rating process and the criteria that raters use as well (Chalhoub-Deville 1995, Connor-Linton 1995, Milanovic et al 1996, Weigle 1998, Shi 2001, Tarnanen 2002). Some projects regard the assessment of a first-language (L1) (Connor-Linton 1995, Schoonen et al 1997, Berge 1996) while others deal with the assessment of second-language performance (Shohamy 1981, 1983, Cumming 1990, Hadden 1991, Halleck 1992, 1995, Chalhoub-Deville 1995, Weigle 1994, 1998, Bachman et al 1995, Lumley and McNamara 1995, Brown 1995, Lynch and McNamara 1998, Shi 2001, Tarnanen 2002). Research on the effect of the rater-variable is conducted in relation to oral performance (Shohamy 1981,

1983, Sieloff Magnan 1987, Hadden 1991, Halleck 1992, Bachman et al 1995, Brown 1995, Chalhoub-Deville 1995, Lumley and McNamara 1995, Lynch and McNamara 1998, Upshur and Turner 1999) as well as written performance (Shohamy et al 1992, Weigle 1994, 1998, 1999, Connor-Linton 1995, Schoonen et al 1997, Berge 1996, Shi 2001, Tarnanen 2002).

Several studies focus on differences between rater groups, and these differences are of several kinds: some highlight the difference between native speaking and non-native speaking raters (Brown 1995, Connor-Linton 1995, Hill 1996, Shi 2001), others on the differences between teachers and non-teacher's evaluation (Hadden 1991, Chalhoub-Deville 1995, Schoonen et al 1997), and even others on the differences between trained and experienced raters on the one hand and inexperienced raters on the other (Cumming 1990, Wigglesworth 1993, Lumley and McNamara 1995, Weigle 1994, 1998, 2000).

Only very few use a before-after design in studying the effect of rater training (Weigle 1994, Lumley and McNamara 1995). The methods of analysis are also diverse: inter-rater reliability is sometimes analysed by using simple methods from classical test theory, such as Cronbach's Alpha (Shohamy 1983, Shi 2001), Pearson Product- Moment Correlation (Hasselgren 1996) or Spearman Brown (Berge 1996), while the most frequently used methods of analysis in recent studies are generalisability theory and MFR (Brown 1995, Hill 1996, Wigglesworth 1993, Lumley and McNamara 1995, Weigle 1998, 2000).

For an overview to be meaningful, some sort of grouping is necessary. Yet, for studies so diverse in focus as well as in design, methods and results, any categorisation will necessarily be a simplification. I have chosen to group the studies according to their research questions and overall research focus. Whether they focus on the assessment of first or second-language performance, whether the skill is writing or speaking, as well as the design or method of analysis are considered subordinate to this.

Taking research focus as a point of departure, then, an overall distinction may be drawn between studies focusing on the effect of the rater variable on reliability, on the one hand, and those focusing on its effect on construct validity and the criteria raters use, on the other. The dichotomy will to some extent overlap with that of early and recent research literature on this topic. Prior to the 1990s, research focusing on the effect of the rater variable on reliability alone seems to dominate. After the 90s, however, the vast majority of rater-related research projects focused on the rating process and the criteria upon which raters base their scores. Hence, rater-related research prior to the 90s may be labelled *product oriented*, while most research of the last decade may be labelled *process oriented* (Vaughan 1991).

7.4.1 Product-oriented research: Studies of the rater effect on reliability.

One example of product oriented rater-related research is Shohamy (1983). Shohamy investigated both intra- and inter-rater reliabilities of an oral interview in Hebrew. She found very high reliability estimates (between $\alpha = .93$ and $\alpha = .99$), calculated by Chronbach's Alpha. The number of raters was limited, though. Only three raters, the investigator included, participated in the study and 100 candidates were evaluated. Shohamy appeals for similar studies where the effect of rater training on reliability is investigated.

In a later project, Shohamy et al (1992) investigated the effect of raters' background (teachers versus non-teachers) and rater training on reliability of a written test. Their results show a positive effect of rater training on reliability, while raters' background did not seem to affect the scores: non-teachers were as reliable in their ratings as teachers. Even though Shohamy et al. did not look into the criteria raters used, they acknowledge the fact that raters may have agreed on the wrong grounds, and that this affects the validity of test scores. They therefore call for similar studies focusing on the criteria raters use.

Another early product-oriented study is presented in Sieloff Magnan (1987). This study investigated the effect of rater training and experience on inter-rater reliability by comparing the scores of academic testers in training (academic trainees) and a master tester using the ACTFL scale. Her findings suggest that trainees are capable of assessing oral performance as reliably as the master tester when basing their scores on the ACTFL scale.

7.4.2 Process-oriented research: studies of the rater effect on construct validity

Research after the 1990s tends to focus on the rater effect on validity as well as on reliability, it is process-oriented more than product oriented, and it is based on a fundamental acknowledgement that agreement about the scores may very well mask disagreement about the criteria, as discussed in Section 7.2.2.

Rater-related research within this tradition may again be divided into subgroups according to the main focus of the investigation and four groups may be distinguished: studies focusing on differences in the evaluation between

1. native speakers (NS) and non-native speakers (NNS)
2. teachers and non-teachers
3. raters with and without rater training
4. studies which investigate rater behaviour based on the ratings of one rater group

7.4.2.1 Native speakers versus non-native speakers

During the last decade, several studies have been conducted comparing the scores given and the criteria used by native and non-native speaking raters. This may have its origin in a practical concern: English is the foreign language for millions of people around the world, and it would be practical and economic if they could be assessed and scored by non-native speakers of their home-countries instead of native speakers of English. In addition, it is of course an interesting theoretical question whether non-native speaking raters focus on different aspects of performance than native speakers.

Brown (1995) used MFR in a study of the effect of raters' linguistic and occupational background on the scores on an oral test of Japanese. She also investigated the criteria used by the distinct rater groups. The results of her study suggest that while there are only minor differences in the scores awarded by native and non-native speaking raters with varying occupational backgrounds (guides, teachers of Japanese, and both) there are significant differences between the groups as to the criteria they use (Brown 1995: 3). Similar results were found by Connor-Linton (1995). His results show only minor differences in the final scores of written EFL essays granted by native speaking (American English) and non-native speaking (Japanese) teachers of EFL. Nevertheless, the two groups of raters focused on different aspects of the written performance. Connor-Linton concluded that "quantitative similarities in ratings may mask significant differences in the reasons for those ratings" (Connor-Linton 1995:99).

An analogous study was conducted by Hill (1996). She used MFR/ FACETS to investigate reliability and construct validity of an EFL test of writing when scored by NNS raters (Indonesian) and NS raters (Australian English) respectively. Her overall results suggest that NNS are equitable raters to NS. There were only minor differences between the two groups: contrary to previous research, Hill found that the NS were significantly harsher than the NNS raters, however. She also found that the NNS raters were more in agreement with each other about the scores, and finally that the two rater groups used the rating scale in similar ways.

In a recent study, Shi (2000) compared the reliability and construct validity of NS and NNS EFL teachers' ratings of the EFL writings of Chinese university students (Shi 2001). Shi focused on rater-agreement as well as on the criteria used by raters in their holistic scoring. She emphasised the importance of letting raters score without any predetermined evaluation criteria, to find out how raters defined the criteria themselves. The results of her findings suggest that NS were more reliable raters than NNS raters, with Cronbach's Coefficients of $\alpha = .88$ and $\alpha = .71$ respectively. At the same time NS used a greater range of the grade scale in their scores than the NNS. As to

the criteria used by the different rater groups, Shi's findings are consonant with those of Brown (1991) and Connor-Linton (1995) suggesting that they based their scores on somewhat different criteria.

7.4.2.2. Teachers versus non-teachers

In most cases, learner language is assessed by language teachers, with or without rater training. It is a likely assumption that through their profession, language teachers gain an increased tolerance for language variation, and one would assume them to understand learner language better than people who are less familiar with foreign accents. At the same time the teachers' job is to improve the language of their students. One could therefore also expect them to be relatively focused on formal aspects of learner language. In any case, it is likely that their daily contact with foreign accents will affect their perception of learner language in one way or another. Several studies are dedicated to the question of differences between teachers and non-teachers' evaluation of learner language.

Hadden (1991) used factor analysis in an investigation of ESL teachers' and non-teachers' reactions to ESL oral communication in a search for qualitative differences between the groups. She found significant differences in the evaluation of non-native speech by ESL teachers and non-teachers: non-teachers tended to interpret a speaker's linguistic ability as interrelated with comprehensibility. The teachers on the other hand evaluated these traits as separate from each other. Hadden also found that the ESL teachers were more critical in their judgement of linguistic ability than the non-teacher group

Chalhoub-Deville (1995) studied the effect of test-method and raters on scores of an oral test of Arabic as a foreign language (AFL). She compared the ratings of one group of AFL teachers living in the USA, with two groups of non-teaching native speakers of Arabic, one of which live in the USA, and the other in Lebanon. As in the study by Hadden referred to above, Chalhoub-Deville found significant differences between the three rater groups. However, as to the criteria used, her findings differ from those of Hadden: Chalhoub-Deville found that the group of teachers focused more on creativity in presenting information, while the non-teaching group residing in Libanon focused almost exclusively on grammar and pronunciation. Hence, the non-teaching group was more preoccupied by formal aspects while the teacher group emphasised more communicative aspects of performance to a greater extent.

Schoonen et al (1997) examined inter-rater and intra-rater reliability, and to some extent the validity, of teachers' and non-teachers' scores of a Dutch writing test for 12 year olds. The

number of raters involved was however limited: only three teachers and four non-teachers were investigated. In accordance with Shi (2001) the results show that teachers used a larger range of scores than the non-teachers. The study shows that rater-reliability is affected by the criteria or rating task raters are to perform: both groups were reliable raters in rating *content,* while the teachers were more reliable in rating *usage*.

7.4.2.3 Raters with and without rater training.

As discussed in Section 7.3 there have been relatively few research studies on the effect of rater training on test scores, considering the widespread use of this procedure in professional language testing. One way of investigating this question is by comparing the scores given by trained raters and lay persons respectively.

Cumming (1990) compares ratings of experienced and inexperienced raters with a particular focus on the decision-making behaviour of each group. Cumming is interested in gathering information about the rater groups' ability to distinguish students' writing expertise and second language proficiency. The results are in favour of the expert SP-raters:

> Overall, expert teachers appear to have a much fuller mental representation of "the problem" of evaluating student compositions, using a large number of very divers criteria, self-control strategies, and knowledge sources to read and judge students' texts. Novice teachers tend to evaluate compositions with only a few of these component skills and criteria […] (Cumming 1990:43.)

Wigglesworth (1993) focused on the effect of rater training on intra- and inter-rater reliability of oral tests. She used MFR to conduct bias-analysis whose purpose was to reveal whether individual raters were consequently harsher in relation to some test types and criteria than for others. The results of the bias analysis were then communicated to the raters, and Wigglesworth aimed to shed light on the effect of this individual feed-back on test scores. Her findings suggest that raters are responsive to feedback, and that information of their individual biases may improve their ratings and should therefore be part of standard rater training procedures.

A similar study was conducted by Lumley and McNamara (1995). They too were interested in highlighting the effect of rater training for intra-rater reliability and rater bias of an oral language test. Of particular interest in their study was the question of the stability of rater characteristics over time. They raised an interesting question: "If a rater's characteristics are successfully modified by training, are these changes stable over time, or does the rater revert to old habits?" (Lumley and McNamara 1995:59), which of course relates to another important questions in relation to rater training: "How often do raters need to be retrained?" (ibid). One important implication of their study is that rater training cannot eliminate variation in rater

harshness, and that FACETS should be used to correct for this kind of rater variability (inter-rater reliability). Rater training does not have a permanent effect on test scores and needs to be repeated frequently.

In her PhD dissertation Weigle (1994) used a before-after design in a study of the effect of rater training on test scores of an ESL test of writing[6]. Weigle's research interest was not limited to the effect of rater training on reliability, because, as she states: "Equally importantly, to what extent does rater training function to bring raters into agreement about the definition of the ability which the test is intended to measure?" (Weigle 1998:265). Weigle compared scores given by inexperienced and experienced raters before and after rater training. The results of the reliability study imply that inexperienced raters are more severe, less consistent and finally more extreme in their ratings than experienced raters. These differences in severity between groups are reduced, but not totally eliminated, after rater training. The results of her qualitative study focusing on construct validity are positive, implying that rater training helps raters to understand and apply the rating criteria in the intended way.

In a later article, Weigle (2000) used the data-set of the 1994 study in an investigation of the interaction of raters and prompts for an ESL writing test. Again she used a combination of quantitative (MFR) and qualitative approaches (raters' think aloud protocols). Her findings imply that inexperienced raters are more severe than experienced raters for one task, but not for the other, and that differences between the groups were eliminated after training. She used qualitative data to investigate the reason for these differences, and found that it was due to the ease with which the rating scale could be applied to the distinct prompts.

7.4.2.4 One group of raters

In contrast to studies comparing the ratings of different rater groups, there are investigations based on one rater group in isolation. Vaughan (1991) focused on the holistic scoring of written essays of trained raters. She found that despite similar training, raters focus on different elements in the essays and that they may have different approaches to assessment of writing. She however warns against generalising the results, as the number of raters is limited.

In a qualitative study of rater behaviour, Halleck (1992) investigated the relative contribution of sentence level grammar and communicative factors on scores of an oral test. Raters were trained in using the ACTFL guidelines as a basis for their judgements. Her main results show that raters were "primarily concerned with communicative strategies rather than with the grammatical accuracy of the interviews" (Halleck 1992:228).

[6] The project is also presented in the 1998 article.

In one of the few Norwegian studies of rater behaviour Berge (1996) investigated the ratings of a written L1 exam. Berge looked at experienced raters' different use of the grade scale in a study of reliability, as well as the underlying reasons for these scores, or the "text norms" in his terminology. The results of the reliability study indicate that raters do not agree about the scores, and that it is particularly hard to agree about which scores to assign to performances in the middle of the scale. Berge did not attempt to measure the criteria upon which raters base their scores directly, rather he gathered different kinds of qualitative data about raters' attitudes towards writing, Norwegian as a school subject and assessment. His results suggest that raters' opinions about what constitutes a good text vary.

Tarnanen (2002) investigates the relation between raters' scores and their reasons for those scores by using a combination of quantitative and qualitative data. The informants of the study are 17 teachers of Finnish as a second language (F2) and they rate 247 written F2 performances. Interestingly, Tarnanen compares scores given by raters when scoring impressionistically and scale-based (as in the present study). Tarnanen finds higher degrees of reliability when raters use a rating scale. The results of the impressionistic scoring are also relatively high, which Tarnanen explains by F2 teachers focus on grammatical traits in their impressionistic scoring.

Summing up some of the main findings of the rather diverse field of rater-related research, it is fair to say that there has been an overall shift in focus from reliability to validity issues during the last decade. This has to do with the introduction of MFR and FACETS which makes it possible to correct for differences in severity between raters. It is however still important within this framework that each rater be internally consistent. Research has shown that different rater groups, such as lay persons, non-native speakers, teachers and trained raters score SL performance differently: they are not equally severe, and they focus on distinct aspects of performance. Most studies find inexperienced raters to be more severe than experienced raters. Rater training may reduce but not eliminate variability due to the rater variable. It is shown to have a positive effect on intra-rater reliability, but be of only limited value to inter-rater reliability: that is, while it seems to be impossible to make raters equally severe, it is possible to make them internally consistent. Studies also show that different rater groups focus on different aspects of performance, but the results are inconsistent as to which groups focus on which aspects. Rather training is found to have a positive effect on construct validity, as it enables raters to understand and use the rating scale appropriately. In the next chapter, the present research project will be placed on the map of the already existing research presented above.

CHAPTER 8: METHOD

Chapter 8 represents a shift of focus from theoretical to empirical issues, presenting research questions and hypotheses, design, participants, data and methods for analysis of the data.

8.1 Research questions (RQ) and hypotheses (H)

The project has been guided by an overall research interest in the value of two highly recommended and frequently used procedures in oral language testing: the use of an explicit rating scale and the training of raters. The main purpose of this study has been to shed light on the effect of these procedures on inter-rater reliability as well as on construct validity of test scores and the main research questions are:

> RQ1: Does the use of trained raters and a rating scale produce raters who are more in agreement about the scores they give, that is, do these procedures have a positive effect on inter-rater reliability?
>
> RQ2: Does the use of trained raters and a rating scale produce raters who are more in agreement with the test constructors about the underlying construct of the tests as specified in the rating scale, in other words, do these procedures have a positive effect on construct validity?

From these principal research questions, four hypotheses are deduced, two of which regard the effect of rating scale and rater training on inter-rater reliability, and two regarding their effect on construct validity:

- H1: **Training of raters** affects **reliability** of scores positively; trained raters show higher inter-rater reliability than untrained raters when scoring both with and without rating scales.

- H2: The use of an explicit **rating scale (NORS)** affects **reliability** of scores positively; inter-rater reliability of scores is higher when raters use a rating scale (the NORS) as opposed to when they score impressionistically. The effect of a rating scale is positive for raters with and without rater training, yet the effect is greatest for the groups of untrained raters (naïve NS and N2-teachers).

- H3: **Training of raters** affects **construct validity** (defined as the match between the criteria of the scale and those of the raters) positively: there is a greater match between the criteria of the NORS and those of the trained raters than between the NORS and the criteria used by other rater groups.

- H4: The use of an explicit **rating scale (**NORS) affects **construct validity** (as defined in H3) positively. There is a greater match between the criteria of the NORS and those of

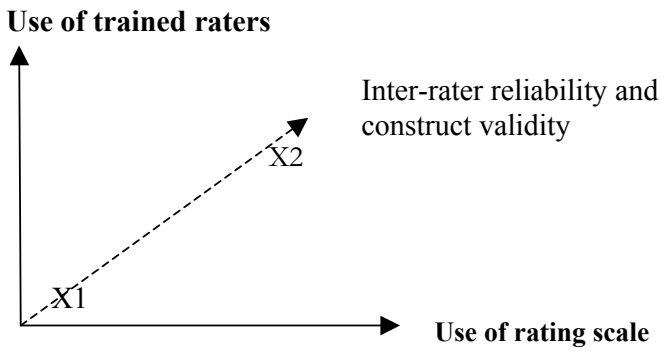
the raters when raters base their scores on the NORS than when scoring impressionistically.

H1 would gain support if trained raters show higher inter-rater reliability than the other rater groups, when scoring holistically as well as with the rating scale (NORS). H2 will be sustained if the results show higher estimates of inter-rater reliability when raters use rating scales as opposed to when they score on purely impressionistic and subjective grounds. This should be true for all rater groups, yet the differences between the scoring methods should be largest for those raters who are unfamiliar with the NORS.

H3 would be supported if there were a greater match between the criteria of the NORS and those of trained and experienced raters, than between those of the NORS and those of raters without rater training. H4 is supported if raters of all groups score more in accordance with the underlying construct explicit in the scale when they have the NORS to hand, than when scoring impressionistically. Yet, the differences between the two scoring methods should be greater for raters who have never seen the NORS before, than for raters who have been trained in using it since trained raters are assumed to have internalised the rating criteria of the NORS and therefore will tend to base their scores on it under all circumstances.

The assumptions of the four hypotheses seen in relation to each other is presented graphically in Figure 11 below:

Figure 11 The predictions of hypotheses H1 – H4.



X1 represents naïve NS when scoring impressionistically, and X2 experienced raters when using the NORS. These combinations of experience and use of rating scales should predict maximum and minimum estimates of reliability and validity if the hypotheses were supported by the data.

According to Popper (1959), a hypothesis can never be confirmed once and for all. Rather researchers should formulate hypotheses, which make precise assumptions and then try to falsify them empirically. If the hypotheses are not falsified by the data presented in Chapters 9 and 10, this supports the claim that the use of rating scales and trained raters are worthwhile in an effort to enhance inter-rater reliability as well as construct validity of test scores.

In Chapter 7 theoretical arguments were presented for the assumed effect of the rating procedure on construct validity. It was claimed that while the subjective scoring of performance based tests was assumed to affect only reliability of scores in traditional language testing, there is consensus in the field of modern language testing that it also affects the construct validity. In modern test literature there is a repeated call for studies investigating the effect of the rating process on construct validity (Vaughan 1991, Shohamy et al 1992, Connor-Linton 1995, Shi 2001). To argue why H3 and H4 are interesting and relevant would therefore be like stating the obvious. Rather, in modern language testing, the hypotheses that need some kind of justification, are H1 and H2. In Chapter 4 and 7 it was argued that inter-rater reliability is considered relatively uninteresting within the framework of IRT, which has come to dominate the research field. Some variation between raters is considered natural and inevitable, and by using MFR, it is possible to correct for it statistically. So why spend time and research effort on it? I will make an attempt to explain by referring to practical as well as theoretical reasons. In Norwegian society a certain degree of rater-agreement is necessary to assure a fair assessment. Even though MFR is a theoretical option, it is not well-known in the testing circles, and hence not used by test constructors. In the school system, at universities and university colleges as well as in the proficiency testing of immigrants, two raters are normally used and the scores they set are the final scores granted. Further analysis or correction of the scores given is, as far as I am concerned, not undertaken in any systematic way. In the real world, then, test takers are entirely dependent on the assessment of the pair of raters that happens to evaluate them. Test constructor should be able to demonstrate that candidates are scored similarly across raters and test administrations, a point also made by Hill: "[…] it is unlikely that, when the test is in use, FACETS will be available to compensate for the inevitable differences in rater harshness (Hill 96: 282).

Another reason why I chose to focus on inter-rater reliability in my study is that it is an explicit aim of the rater training undertaken at Norsk språktest (1998b). Norsk språktest develops rating criteria and trains the raters in applying these, in order to heighten the convergence between raters about the scores they give. An investigation into the effect of these time consuming and costly procedures therefore seems justifiable.

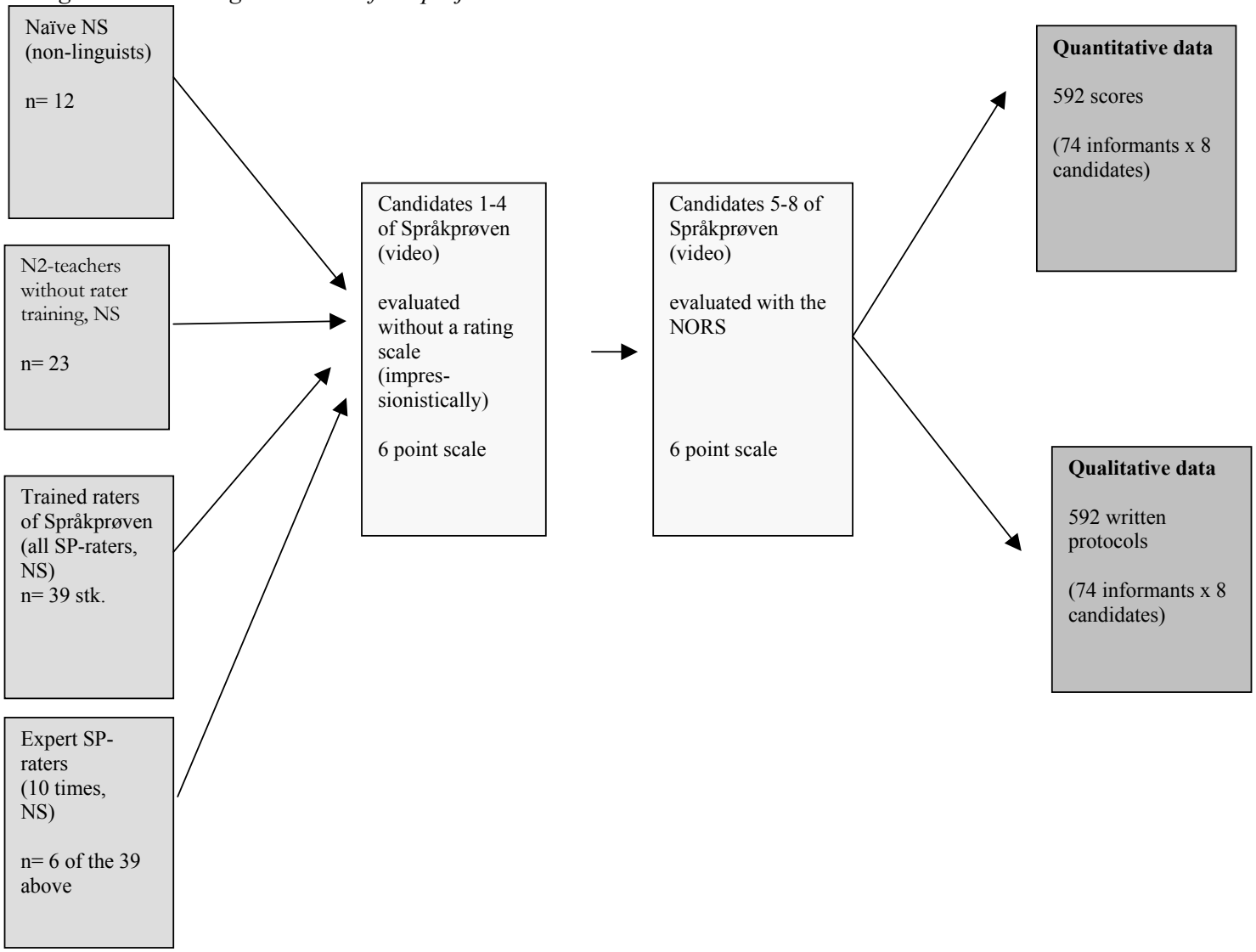
In addition to these prosaic reasons, there are weighty theoretical arguments for focusing on rater agreement within a modern language test framework. As Bachman argues, inter-rater reliability is affected by "inconsistencies in the criteria used to rate and in the way in which these criteria are applied" (Bachman 1990:180). If raters disagree about the scores they give, this may very well be due to the fact that they base their scores on different criteria. This of course, is a serious problem for the test, as it affects, not only the reliability, but the very construct validity of scores. Consequently, test researchers should keep focusing on rater agreement, even though it is technically possible to correct for discrepancies between raters using MFR and FACETS. Along these lines, Cushing Weigle (1998) warns against a disinterest in inter-rater reliability, as she claims it may have negative side-effects for construct validity:

> […] a de-emphasis on inter-rater agreement may have implications for the construct validity of the test if it draws attention away from getting raters to agree on a definition of the ability being measured by the test. In other words, if internal consistency is considered the most important benefit of training, and differences in rater severity are compensated for mathematically, a thorough understanding of the intended grading criteria may no longer be a central aspect of the training process. Raters may in fact learn to interpret the scoring rubric in idiosyncratic ways, each of which may be consistent in itself but which may not have anything to do with the construct […] as defined by the test writers (Cushing Weigle 1998: 265).

8.2 Design and data

The design and the data used in this thesis are compound and may perhaps best be presented graphically:

Figure 12 The design and data of the project.



As the figure shows, the empirical part of the project has three main components: the participants or informants described in the boxes to the left, the scoring of eight N2 candidates described in the two boxes in the middle of the figure, and finally the empirical data, which consist of a combination of qualitative and quantitative data presented in the boxes to the right. We shall look at the different components in turn.

8.2.1 Informants

In this thesis a claim is made that raters' background, training and interaction with the rating scale are of crucial importance to reliability as well as validity of test scores. In order to make generalisations of the results, it is of great concern that the sample of informants be representative of the population to which one wishes to generalise. One way of assuring a representative sample is by random selection. But for studies such as this one, where the participants have to volunteer to take part in it, it is difficult to conduct a true random selection since informants' willingness to participate in the study, may represent a bias: one could expect to get an over-representation of informants who like the task they are asked to perform, while those who feel uncomfortable with it, or simply find it uninteresting, will be more reluctant to take part in the study. The trained raters of Språkprøven might get a feeling that their qualifications as raters are being put on trial, and those who find the rating task difficult will naturally hesitate to volunteer. This problem in relation to the use of human informants can hardly be completely overcome. Nevertheless, there are means of coping with it. Hatch and Lazaraton (1991: 43) stress the importance of creating a pool of volunteers from which the random selection can be made. Still, the problem remains that all informants have volunteered, and probably for the reasons sketched above. Another, in my opinion, better type of remedial action is to pay the participants a fee. Even though the problem sketched above remains, the fee may attract other informants too. The higher the fee, the more important the effect of it will be, of course. The informants of my study were offered a fee of 250 NKr per hour. In addition, the sample has been selected at random from a larger pool of volunteers as recommended by Hatch and Lazaraton. This was done as I first contacted schools of N2 for adults asking for a list of teachers who would like to participate, I then awarded a number to each of them and performed a random selection based on a table of random digits (Butler 1985:169).

From a micro-perspective I aim at generalising the results of the study to the population of existing and potential raters of the oral part of Språkprøven, in order to learn whether the training and the use of a rating scale (NORS) actually make them better raters. From a larger

perspective, I wish to generalise to any situation where raters interact with rating scales in a test situation. The research questions relate to every situation where test constructors make use of trained raters and explicit rating scales in an attempt to heighten reliability and validity of assessment involving qualitative judgements.

The final number of informants in the study is 74[7]. They are sub-categorised into three main groups: Group 1 consists of 12 naïve NS, or naïve native speakers (naïve NS). None of them have experience in N2-teaching, nor have they any knowledge of or experience with Språkprøven or the NORS. However, they all have higher education, that is between five and eight years at university level, in an attempt to match this group with the groups of teachers and raters. Group 2 consists of 23 N2-teachers of the state-run educational program for foreign adults. As with Group 1, participants of this group are native speakers of Norwegian and they have higher education from either university or university college for schoolteachers. Being teachers of N2 some kind of testing and evaluation is part of their job: teachers will inevitably have to reflect on their pupils' progression, they have to gain insight into what they know, what they have learnt and what they need to focus more on in class. Informants of Group 2 have not, however, received rater training in the use of the NORS in rating the Språkprøven, nor have they been used as raters of this test. And finally, Group 3 consists of 39[8] trained and experienced raters of Språkprøven (all SP-raters). They have been trained in using the NORS and have between two and ten times experience in rating the oral part of Språkprøven applying this scale. Like the participants of Group 1 and 2, they are all native speakers of Norwegian and they all have higher education. As with the informants of Group 2, raters of Språkprøven are also N2-teachers. For some of the analyses, Group 3 has been further divided: as mentioned, the raters of Språkprøven have varying degrees of experience, so do the raters participating in the study. It seemed reasonable to investigate potential differences between raters with limited experience and those with extensive experience (expert SP-raters). This group consists of raters of Språkprøven who have rated on more than ten occasions, that is for at least three years.

The sample could have been further stratified by including only raters of a certain age, sex, years of teaching practice, educational background, years of experience with the NORS etc. I have chosen not to do this due to two factors: a highly stratified sample would limit the degree of

[7] The number of informants of the quantitative study is 73 because there was a missing score of one of the candidates by one of the raters. This informant is however included in the qualitative study, and the number of informants is here 74.

[8] The one informant who was excluded from the quantitative study because of a missing value, belongs to this group. The number of the group all SP-raters is therefore different for the two datatypes: 38 for the quantitative and 39 for the qualitative study.

general application of the results. Moreover, it would limit the pool of suitable raters. Norway is a small language community, and the number of possible informants is not large. It is simply hard to come up with a sufficiently large number of native-speaking informants who are N2-teachers of adults, between 30 and 40 years, female (or male), have the same kind of educational background, have the same years of teaching practice and fit the specification for training/experience with NORS. The participants of the study have however been asked to fill in a form supplying personal data, allowing me to investigate the effect of these variables on their ratings. This information was collected through a set of questions that raters answered in writing. In addition to questions about age, sex, and native language, they were asked about their formal education, their experience as N2-teachers and their experience as raters of oral tests other than Språkprøven. Finally, they were asked questions in relation to Språkprøven and the NORS, such as how many times they have received rater-training for Språkprøven as well as how many times they have scored the oral part of this test. This information allows a set of interesting research questions to be raised. One could look at differences in ratings by men and women, younger and older raters, with long or short education, etc. In this thesis, the information has been used mainly as a basis for dividing the informants into the three groups discussed above according to their teacher experience on the one hand, and rater training and experience on the other.

8.2.2 Procedure for data collection

The two boxes in the middle of Figure 12 represent the data-sampling part of the thesis. It contains the following: the video taped oral performance of eight N2-learning candidates of Språkprøven and the holistic versus NORS-based scorings of all 73 informants.

Eight *candidates* provided the speech samples for the study. Norsk språktest has a large number of oral test tasks performed by different test takers on video, and I was given permission to use this material. A sub sample of the performances had been used in the production of a video for rater training. Those samples were of course excluded from my study. As far as possible candidates who were not on the training video were selected, but experienced raters may have seen some of them performing other tasks. Still, I presume the effect of this to be minimal as a candidate may perform better on one task than on another, and therefore receive distinct scores on different tasks. Additionally, I assume the possibility that raters remember the scores of each candidate on the video a year after the last training section to be minimal[9]. The eight speech samples were not randomly selected, but carefully chosen in accordance with certain criteria considered relevant for the investigation: there should be a spread of language ability: ideally, all

six levels of the NORS should be represented. I however failed to find candidates who were awarded the top and bottom scores, 1 and 6, by the expert rater team. There should be a similar spread of proficiency between candidates of both groups, those scored impressionistically and those scored by raters using the NORS. It is easier to agree on extreme candidates (top and bottom of the scale) than about performances in the middle of the scale. It could therefore be a source of error if candidates of one group were lumped together in the middle of the scale (3-4) while candidates of the other group represented extreme scores at the top and bottom of the scale. Different learner profiles should be represented, i.e. that some candidates are verbally active, take risks and try to explain relatively complicated thoughts even though this may lead to more grammatical errors, while others focus more on formal traits of their performance and prefer to avoid difficult situations by answering the examiners' questions as briefly as possible. The examiner should act discretely and appropriately so that the raters' focus stays with the performance of the test takers and not with that of the examiners. A final criterion for selection of candidates is technical: sound and image should be clear and sharp.

The candidates are of different age, sex, nationality, ability level, and personality. A short presentation of the candidates therefore seems justifiable. Their names are, of course, fictitious: *Ute* is a German woman in her late 20s. She is judged by the team of professional raters to be at a proficiency level three on the NORS scale. She is relatively shy and depends to a great extent on the collaboration of the examiner. *Ivan* is a Russian man in his early 20s. He is a student in Norway, as he was in his home country, and claims to have a lot of Norwegian friends. He is rather talkative and extrovert and seems to feel quite at ease in the test situation. The rater board placed him at five on the six-point scale. *Hawa* is a woman in her mid 30s working as a hair-dresser in Norway, as she did in her home country Gambia. Despite her limited language resources, she interacts quite actively with the examiner. The board granted her a grade three on the six- point scale. The last candidate to be scored impressionistically is *Beriz,* a Bosnian man in his early 40s. He talks a lot during the test, yet there is no real interaction. Rather there is a monologue on the part of the candidate. The board has given him a five. The next four candidates are scored by the raters using the NORS. The first one is *Yasin*, who is a man originally from Iraq, and who is now in his early 50s. His Norwegian skills are poor, and the board placed him on level two. *Kate* is a student from the USA. She is in her early 20s. She is rather silent, and seems to be the kind of person who is in fear of making mistakes, who prefers to say as little as possible. In fact, some of the informants commented that she produced too little language for a fair judgement of her skills. The board however granted her a four. The next

[9] The last time NLT gave training course for oral raters before the data collection of the study.

candidate, *George*, is also from the USA. He is obviously nervous, and totally dependent on the help of his examiner. The language he produces to a large extent restricted to words and sentence fragments, yet similarities between his first language and the target language probably helps him getting his message across. The board gave him a three. Finally, *Karlo* is from Bosnia and in his late 20s. He is the one who is granted the best score of the eight candidates. He was finally awarded a five, but several of the raters of the board would consider giving him the top score six. Karlo is relaxed and talkative, and carries out a conversation with his examiner very similar to an everyday conversation between peers[10]. The spread of proficiency levels of candidates of the two groups are: for the group that is scored impressionistically 3, 3, 5, 5 and for the group scored by raters using the NORS: 2, 3, 4, 5 (6).

Different scoring methods were discussed in detail in Chapter 6 and will only be dealt with briefly here. In order to investigate the effect of the use of rating scales on reliability and validity, I compare scores given by the same raters when they score with and without an explicit scale. For the impressionistic scoring, raters were not given any rating criteria or guide whatsoever but asked simply to place the candidate on a six-point scale and give the reasons for the scores they set. The next four candidates were scored by the same raters using the NORS, again on a six-point scale. With reference to the categorisation of rating scales given in Chapter 6, it is evident that the NORS does not fit into either holistic or multiple trait scales: it contains a specification of different rating criteria treated independently of one another as is common in a MTS, yet raters only set one score on the whole performance, as is common in a holistic rating scale[11]. The criteria specified in the NORS are: communication and verbal initiative as the principal criterion, vocabulary, grammar and pronunciation as the subordinate criteria. Level descriptions are given on three main levels, 1-2, 3-4 and 5-6. Formal correctness is linked to intelligibility, the ability to get a message across and to whether or not the mistakes lead to misunderstandings. It is based on a communicative view of language. For more details, see Chapter 6.

The procedure for *data collection* is of relevance for the results. Since the data were not collected at one point of time, it was necessary to standardise the procedure. This means in part that informants were given exactly the same piece of information before seeing the candidate performances. This information consisted of the following:

[10] Transcriptions of the candidates' performances are attached in the Appendices.
[11] This is different in the revised versions of Språkprøven, where raters set separate scores for the different criteria as common in MTS.The separate scores are added and serves as a basis for giving one of the following three grades: ikke bestått/ bestått/ godt bestått (fail/ pass/ pass with excellence.

- Short description of the intermediate proficiency level
- Short explanation of how to use the six point scale: informants were made aware that they were not allowed to give in-between scores, but had to decide upon one score only
- They were told not to focus on the examiner's role or the test tasks, but to concentrate on the test-taker's performance
- Short explanation of how to interpret the personal data questionnaire

The standardisation of the sampling procedure also means that raters were asked to score the candidates in the same order: they first scored candidates 1- 4 impressionistically, that is without a rating scale, and placed them on a six point grade scale. They had no verbal specifications to guide them except from the anchors of the top and bottom levels of the scales: "very good" and "very poor". Thereafter they scored candidates 5- 8 using the NORS. They were not given any explanations for the descriptors of the NORS in order to avoid untrained raters from getting trained during the sampling process. After each score they were asked to answer additional questions in relation to their scores: "Why did you not give a higher score?" and "Why did you not give a lower score?", a procedure also used by Halleck (1992). They were made aware that there were no constraints imposed on these explanations, and I did not guide them in any way if they asked questions in relation to the ratings. This, however, rarely happened: on the contrary, it seemed that the initial information was sufficiently explicit for the informants to perform their task without further questions as to what they were expected to do[12]. The duration of each sampling session was approximately two hours. The introductory information part took about 10 minutes, each video recorded task lasted for about 5 minutes (total: 40 min.), the raters were given 7 minutes to score and give an explanation for their scores (total: almost 60 min.) and finally 10 minutes to fill in the personal data scheme. Handing out and collecting sheets from the informants took another 10-15 minutes. Data were collected in Bergen, Oslo and Kristiansand at the state run programs for the teaching of Norwegian as a second language for adults at the schools of Nygård, Rosenhof and Lahelle respectively. The data-collection was a time consuming enterprise, both for the participants and for the researcher, something which has limited the size of the data sample. However, compared to similar studies of rater reliability and rater behaviour, the number of raters involved in the present study is comparatively large: Schoonen et al (1997) use as few as seven raters, Cumming (1990) uses 12, Lumley and McNamara (1995) 13 and Weigle (1994) 16 raters. Tarnanen uses 17 raters, Shohamy, et al (1992) use 20, Hill (1996) 23, Brown (1995) 33, , Shi (2001) uses 46, Connor-Linton (1995) 55 and Hadden (1991) 57. As far as

[12] A subgroup of the raters were also asked to answer a set of open-ended questions in relation to oral performance, rating, rating criteria and the NORS. These data have not been exploitet in the present study, but may form the basis of a future investigation of raters' attitudes of non-native speak.

I am concerned, the only study that has a larger sample of raters in their study is Chalhoub-Deville (1995) who uses as many as 82 raters in her investigation. However, in several of the studies mentioned above raters are asked to score a larger number of candidates than in the present study, especially in studies focusing on the rating of written essays.

8.2.3 Data and analysis

In order to investigate the construct validity of performance based tests, raters' criteria should be compared with those made explicit in the rating scale, as discussed in Chapter 7. The numeric scores given, do not reveal the underlying criteria of the raters, and quantitative data are therefore insufficient in a study like the present.[13] The study therefore bases its investigation on a combination of quantitative and qualitative data, as recommended by Bachman[14] and practised by Weigle (1994), Connor-Linton (1995), Shi (2001), Tarnanen (2002) and others.

The quantitative data consist of the 74 raters' scores of each of the eight candidates on a six-point scale, that is in total 592 scores. These data were coded in SPSS and related to the background information about raters for statistical analysis. The scores not coded as interval data, since I do not know whether the distance between the scores on the scale are equally distributed. Studies have shown that on a six-point scale, the distance is normally larger between 3 and 4, than between 1 and 2 (McNamara 1996, Linacre 1999):

Actual step structure: 1--2-----3-------4----5-------6
Apparent step structure. 1-----2-----3-----4-----5-----6
(McNamara 1996:125).

Moreover, candidates who score 4 on a scale do not necessarily possess twice as much of whatever is measured as candidates who score 2 (Bachman 1990). These data were therefore coded as ordinal scale data.

The quantitative data were used in an investigation of the effect of rater training and rating scale on inter-rater reliability, i.e. they were used in the testing of H1 and H2 presented in Section 8.1. The inter-rater reliability estimate was calculated by using *Cronbach's Alpha* (= α). Cronbach's Alpha is recommended as one of the methods of classical test theory suitable for calculating the level of agreement between raters of a group. (Crocker and Algina 1986: 138, Shi 2001). In modern test research, it is however more common to analyse rater variance within the frame of IRT and by the use of approaches such as generalisability theory MFR. I have already

[13] The exception is when raters use an analytic or multiple trait scale where they give separate scores for separate traits. This is not the case for the NORS where raters give one final score.

argued in Section 8.1 why I consider inter-rater reliability to be of crucial importance in assuring a fair assessment, at least in Norwegian society where the use of MFR to correct for differences in rater severity is still in the future. For research purposes, the main advantage of MFR is, as I see it, that it allows us to investigate interactions between variables affecting test scores such as rater severity, item difficulty and the underlying ability of candidates (Lumley and McNamara 1995, Lynch and McNamara 1998). This has not been the main purpose of the present study however. Another important advantage, though, is that it detects misfitting raters, that is raters who are not internally consistent or who give scores that differ extremely from those of the other raters. This would without doubt have been an advantage in the present study as well.

I have been asked why I do not use quantitative data in an investigation of raters' ranking of the candidates. Even if raters were not totally agreed about the scores they give each candidate, it is possible that they still rank the candidates in the same order. I have chosen not to focus on this aspect because I consider raters' ranking irrelevant in relation to the test under study here. Språkprøven is a criterion-referenced test i.e. raters place the candidates not according to each other or a norm group, but with reference to proficiency levels on a rating scale (see Chapter 4 for a discussion of NR versus CR). A fair assessment therefore depends on raters' ability to place the candidates on the same level of the scale and not in relation to each other.

The quantitative data yield information about rater severity, and about the way different raters and rater groups assess the distinct candidates. It is used in this project primarily in the study of rater-agreement about the scores, and differences in rater-agreement between different rater groups (naive native speakers, N2 teachers and trained raters of Språkprøven). But another important research question of this project, is the effect of these procedures on construct validity. As mentioned, the quantitative data are not suitable for discovering the underlying criteria of raters, and an additional set of *qualitative data* was therefore collected.

There are mainly two ways of collecting information about the underlying criteria of raters' holistic scores: concurrent think-aloud protocols on the one hand and retrospective written reports on the other. In the *concurrent think- aloud protocols* raters tape- record their thoughts during the rating process and it is assumed to be a reliable way of investigating what goes on in the rater's mind during the rating process. However, analysis of the data requires transcription, which is very time-consuming. For a relatively large number of raters as in this study, think-aloud protocols would therefore be hard to accomplish within the limitations of the project. In addition, it is less suited for oral assessment than for the assessment of essays: raters would have

[14] e-mail communication 1999

to watch and comment on the candidates in separate rooms in order not to disturb each other while taping their comments. An additional disadvantage is mentioned in Lumley (2002). He compares scores given by raters when rating with and without think-aloud protocols. Lumley claims to have found evidence that concurrent think- aloud protocols disturb the rating process: "The think-aloud requirement causes significant disruption to the rating process, and the rating behaviour of some raters is altered by the think-aloud requirement" (Lumley 2002). One should therefore interpret the results of studies using this method with caution, he argues. Think-aloud protocols are used in studies of rater behaviour by Cumming (1990) and Weigle (1994, 1998) in relation to writing.

The other procedure apt to gain insight into raters' judgements, are retrospective *written reports* (WR). Raters listen to an oral speech sample, or read a written text, and thereafter try to articulate in writing the reasons for their scores. The advantage of this method is that it is normally easier to analyse as raters tend to be briefer and more precise in their written explanations than in think aloud protocols. In addition it does not disturb the rating process as the written report is given after the score is set. This method is used by Connor-Linton (1995) and Shi (2001), and it is the one used in the present study as well.

Raters first set a score, and then argue why they selected that particular score. They are not given any restrictions as to which criteria they may use in their explanations. Several researchers investigating raters' criteria emphasise the importance of this procedure. In order to capture the true norms of raters, they should not receive any guidance as to which aspects to focus (Connor-Linton 1995, Shi 2001, Niedzielski and Preston 2000):

> This ensures that the reasons [raters] gave for their ratings are not the result of the researcher's assumptions about what kinds of reasons are important, and therefore can more accurately reflect what was salient […] to each group of raters (Connor-Linton 1995:100).

The obvious problem in relation to the use of WR without any kind of predetermined criteria, is the fact that raters' explanations and varying criteria have to be interpreted and categorised by the researcher afterwards. As does Connor-Linton (1995), I base the categorisation on the most frequently referred criteria by the raters, which were *grammar, pronunciation, vocabulary, fluency, communicative ability, intelligibility, comprehension, initiative, strategies* and *content*. Obviously, if this categorisation were based on a theory of language such as for instance Bachman's model of CLA instead of on folk-linguistic norms, several of these criteria would be grouped together under the same label. The criterion communicative ability could indeed include all the other criteria (in models of CC formal correctness of grammar, vocabulary and pronunciation are an integrated

part, as discussed in Chapter 2), while raters of this study often tend to use it as opposite to formal correctness. Many raters tend to comment on formal mistakes, and then make a positive comment in relation to the candidate's communicative ability, or ability to get the message across, to understand and to be understood. The classification of raters' criteria according to the most frequently mentioned categories was however for the most part surprisingly straightforward: the informants did to a very limited degree refer to unpredictable criteria, rather they used a meta-language in their comments quite acceptable for a linguist. This may be due to the sample of raters: Most of them were linguists (N2-teachers and raters). The informants of the group of lay-persons were all highly educated, probably with knowledge and consciousness about language, even though formally classified as naïve NS. In the few cases where it was too hard to interpret an informant's WR, I choose not to include the comment rather mis-placing it under the wrong criterion.

In the second place, I do however group the criteria in two main categories: formal linguistic traits, on the one hand, and communicative functionality on the other, where *formal linguistic traits* include *grammar, vocabulary* and *pronunciation* and communicative functionality includes the other criteria except fluency and content.

The qualitative data are analysed in two ways: firstly, I calculate the percentage of raters of the different rater groups who use the distinct criteria in their WR. Thereafter I calculate the frequency by which the different rater groups refer to each trait. It is possible that all raters focus on a certain trait but that raters of some groups refer to some traits more frequently than raters of other groups. A combination of percentages of raters and frequency of use may therefore yield more interesting information than any one of the two used in isolation.

Summing up the main content of this chapter, I have four hypotheses for my project, all of which concern the effect of the rater variable on the score of oral language tests. I combine qualitative and quantitative data; H1 and H2 are tested against the quantitative data, while H3 and H4 are tested against the qualitative data. The approach is deductive and hypothesis -testing for both parts of the study.

CHAPTER 9: RESULTS OF THE QUANTITATIVE ANALYSIS

This chapter presents the results of the quantitative analysis in relation to hypotheses H1 and H2 presented in Chapter 8. A more thorough discussion of the results is given in Chapter 11. The aim of the quantitative study is to investigate the effect of two procedures which are assumed to have a positive influence on inter-rater reliability, that is rater training[15] on the one hand, and the use of an explicit rating scale, on the other.

The hypotheses tested on the quantitative data are as follows:

According to the predictions of H1 and H2 naïve NS and N2-teachers without rating experience are assumed to achieve low estimates when scoring without explicit criteria. When using the NORS their scores are expected to be more in agreement with each other. The reliability estimates obtained by the group of experienced raters, however, are expected to surpass the other groups not only when raters use the NORS but when they score impressionistically as well. This is based on the assumption that raters eventually internalise the level descriptors and the proficiency levels of the rating scale, so that they base their scores on the scale even when they do not have the rating scale at hand. As a consequence of these assumptions, experienced raters should obtain higher estimates of IRR and vary less across scoring methods than the other rater groups.

As mentioned in Chapter 8, the inter-rater reliability was calculated by using Cronbach's Alpha. The results are presented in Table 1 below:

[15] Rater training was defined as formal training in rater sessions as well as the informal training of real test evaluation. Trained raters therefore means trained and experienced raters, even if the degree of experience varies.

Table 1 IRR estimates, comparison of groups, naïve NS, N2-teachers, all SP-raters.

Informants	Impressionistic scoring α =	NORS-based scoring α =
Naïve NS (n = 12)	.24	.47
N2-teachers (n = 23)	.12	.40
All SP-raters (n = 38)	.38	.36

The table should be read as follows: Naïve NS obtain reliability estimates of **α***=.24 when scoring impressionistically and .47 when basing their scores on the NORS etc.*

The group that achieves the highest estimates of internal agreement is the group of naïve NS when using the NORS (α = .47), while the poorest estimates of agreement is achieved by the group of N2-teachers when scoring impressionistically (α = .12). All SP-raters show the highest internal agreement when scoring without a rating scale (α = .38) and vary the least from one scoring method to the other, as will be discussed in more detail later. No informant group attains reliability estimates above the recommended .70 limit[16]

H1 is supported for the impressionistic scoring method, but not when raters use the NORS. When scoring impressionistically, all SP-raters do indeed obtain the highest reliability estimate of the groups. When using the NORS, on the other side, the group of all SP-raters actually obtains the lowest estimates of the groups and naïve NS the highest. Consequently, rater training seems to have a positive effect when raters score impressionistically but not when they have an explicit rating scale at hand. H1 is therefore sustained for one scoring method but not for the other. It is however important to notice that the differences between groups are small for the NORS-based scoring.

H2 predicts that raters score more in accordance with each other when using the NORS than when scoring impressionistically. This hypothesis is supported by the data for the groups with no prior rater training, but not for the group of all SP raters. The largest difference in reliability-estimates from one scoring method to the other is seen for the group of N2-teachers (.12 vs. .40). The naïve NS also show a considerable increase in reliability from .24 to .47. The group of all SP

raters, however, shows only a small decrease from .38 to .36. It would probably be right to claim that for this group, there is almost no difference in reliability whether raters score impressionistically or with the NORS.

The most surprising finding of the reliability calculation was the relatively small differences of reliability between the group of all raters of Språkprøven and the groups of informants with no rater training when scoring with the NORS. Even more astonishing perhaps, was the fact that, when using the NORS, the group of all SP-raters was outperformed by both N2-teachers and naïve NS. The lack of internal agreement between raters of this group obviously needs an explanation. In Chapter 11 I search for possible explanation in raters' use of criteria. There might however be a rather pragmatic reason for the results: the group of all SP-raters is, as mentioned in Chapter 8, heterogeneous when it comes to the degree of experience as raters of Språkprøven. Even though all SP-raters have received rater training, their experience from live ratings varies from as few as one or two times up to more than ten times. The reference to raters of Språkprøven above includes all SP-raters regardless of their varying degree of experience. The results analysis would perhaps have been different if only raters with a certain degree of experience were included in the study. Table 2 below presents a comparison of the three groups of informants, but here the group of all SP-raters is replaced by a group of SP-raters with more than 10 times of live rating experience, in the following called expert SP-raters.

Table 2 IRR estimates, comparison of groups, naïve NS, N2-teachers, expert SP-raters.

Informants	Impressionistic/ no rating scale	NORS
	α=	α=
Naïve NS (n = 12)	.24	.47
N2-teachers (n = 23)	.12	.40
Expert SP-raters / more than 10 times (n = 6)	.69	.45

An obvious disadvantage of this approach is the reduction of informants from 73 to 41, and the number of SP-raters from 38 to 6.

[16] It is not unlikely that other reliability estimates could have given higher values for all groups. Since the interest here is on group differences, this is not a major problem. The relation between the groups would most likely be the same.

As apparent in the table the expert group obtains higher degrees of IRR than the group of all SP-raters for both scoring methods. For the impressionistic scoring method there are substantial differences between rater groups. Expert SP-raters' scores are considerably more reliable than those of the other groups and close to the acceptable limit. When scores are based on the NORS, however, there are only minor differences between the groups, and again, the group of expert SP-raters is outperformed by the group of naïve NS. Once more, then, H1 is only partly sustained by the data: rater training seems to make a difference on inter-rater reliability for the impressionistic scoring but not when raters use an explicit rating scale. If we compare reliability estimates across scoring methods, we see an increase of internal agreement for the two groups of raters without rater training (naïve NS and N2-teachers), but the group of expert SP-raters shows a decrease in reliability from impressionistic to NORS-based scoring. For this groups, having the NORS at hand does not have a positive effect on inter-rater reliability. The results will be discussed in detail in Chapter 11.

The main findings of the quantitative investigation were as follows: when all SP-raters are treated as one group and compared to the groups of naïve NS and N2-teachers when scoring impressionistically, the group of all SP-raters shows the highest degree of internal agreement as predicted by H1. Yet, the reliability estimate obtained (.38) is far below acceptable. When raters base their scores on the NORS H1 is falsified by the data: now the group of all SP-raters is outperformed by both N2-teachers and naïve NS. When the group of all SP-raters is replaced by the subgroup of expert SP-raters, there is a notable increase in inter-rater reliability estimates. The group of all SP-raters obtains estimates of .38 when scoring impressionistically and .36 when basing scores on the NORS, while the subgroup of expert SP-raters obtains .69 and .45 respectively. When scoring impressionistically the group of expert SP-raters outperforms the other groups with considerable differences: .69 over .24 (naïve NS) and .12 (N2-teachers), and H1 is again sustained. When scores are based on the NORS on the other hand, the hypothesis is falsified by the data, but there are only minor differences between the groups.

H2 predicts a positive effect of the use of the NORS on inter-rater reliability: raters of all groups should obtain higher estimates of reliability when using the scale than when scoring impressionistically. This hypothesis is sustained by the data for the groups of informants with no rater training or experience with Språkprøven, (naïve NS and, N2-teachers) but not for the two groups of raters of Språkprøven. All SP-raters see a minor decrease from impressionistic to scale-based scoring, while the group of expert SP-raters shows a sharp decrease from .69 to .45. The

results of the reliability study, and particularly the more surprising findings, will be discussed in detail in Chapter 11.

CHAPTER 10: RESULTS OF THE QUALITATIVE INVESTIGATION.

In this chapter the results of the qualitative investigation are presented and related to the effect of rater training and rating scale on construct validity defined as the match between raters' criteria and those of the test constructors specified in the rating scale. An additional purpose of the qualitative study is to search for explanations to the somewhat surprising results of the reliability investigation and this question will be approached in Chapter 11.

H3 and H4 are tested against the qualitative data in Section 10.3. In Section 10.1 and 10.2 I present some differences between the use of criteria, first between rater groups, and second between scoring methods without linking them directly to the hypotheses. This may seem superfluous but it nevertheless felt necessary to illuminate the use of criteria from different angles before approaching the hypotheses. The questions I raise in these first sections are:

- whether different rater groups focus on distinct criteria
- whether raters focus on different criteria when they score impressionistically or NORS-based

It is important to bear in mind that these questions do not relate to the construct validity of test scores. In order to investigate this question, the criteria of the raters should be linked to the criteria of the rating scale in question. This will be done in Section 10.3.

10.1. Do different rater groups focus on distinct criteria?

Firstly, results of both percentages of raters and frequencies of use are reported separately for each rater group. Thereafter the results are combined and discussed in relation to the research question in focus.

Table 3 Ten criteria, naïve NS (n = 12), both scoring methods, percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impressionis-tically	100 %	100 %	92 %	75 %	42 %	75 %	58 %	0 %	0 %	8 %
With NORS	100 %	92 %	100 %	42 %	58 %	75 %	67 %	58 %	8 %	0 %

The table should be read in the following way: When scoring impressionistically (the first row) 100 % of the naïve NS refer to grammar and pronunciation in their WR. 92 % refer to vocabulary etc. The next row refers to their scores, when based on the NORS.

All raters of this group focus on the formal linguistic traits: *grammar* and *pronunciation. Vocabulary* is used by 92 %. Two thirds of the raters of this group stress *fluency* and *intelligibility* and about half of them emphasise *communicative abilities* and the candidates' *comprehension skills*. No raters use *initiative* or *strategies* in their argumentation, and only 8 % stress *content*.

When using the NORS, formal linguistic traits are still used by the majority of raters of this group. Almost all raters focus on *grammar, vocabulary* and *pronunciation*. *Fluency* on the other hand, falls from 75 % to 42 % of the raters. *Initiative*, which was not used by naïve NS when scoring impressionistically is now used by as many as 58 % of the raters[17].

Table 4 Ten criteria, naïve NS (n = 12), both scoring methods, frequencies.

Scoring method	Grammar [18]	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Impressio-nistically	4.08	3.08	3.83	1.33	1.00	1.75	1.42	.00	.08	.08
With NORS	1.00	2.92	4.00	0.75	1.50	2.42	1.42	1.83	.08	.08

The table should be read as follows: Each rater refers to the criterion grammar in average 4.08 times when scoring impressionistically and 1.00 times when scorings are based on the NORS, pronunciation is used in average 3.08 times by each rater when scoring impressionistically, and 2.92 times when scores are based on the NORS, etc.

Grammar is the most frequently used criterion by the naïve NS group when scoring impressionistically. The non-linguists refer to this trait approximately four times on average, followed by *vocabulary* (3.83) and *pronunciation* (3.08) which are both used more than three times by each rater. *Intelligibility* (1.75) is used less than the formal linguistic criteria but more than *communicative ability* (1.00), *fluency* (1.33) and *comprehension* (1.42). *Initiative, strategies* and *content* are hardly used at all by this group of raters when scoring impressionistically.

When using the NORS the most striking difference is that *grammar* is losing ground: G*rammar* is now only used once by the raters of this group. The use of *initiative*, on the other hand, expands from zero to 1.83. The other traits, v*ocabulary*, *pronunciation, fluency, communication, intelligibility, comprehension, strategies* and *content* see no major changes in use from impressionistic to NORS-based assessment.

[17] For a short repetion, the criteria of the NORS are *communication and initiative* as the superior criteria, while *vocabulary, grammar* and *pronunciation* are included, yet as subordinate to communication and intiative.

[18] Even though there are only four candidates to be evaluated through each scoring method the mean frequency may exceed 4.00. This is due to the fact that raters are asked to argue both why they did not give a higher, as well as why they did not give a lower score. A criterion may be used as much as eight times for the impressionistic and eight times for the NORS-based evaluation if a rater refer to the criteron in every explanation given, negative as well as positive. There are a few examples of this in the data.

Table 5 Ten criteria, N2-teachers (n = 23), both scoring methods, percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impressionis-tically	96 %	74 %	96 %	61 %	74 %	74 %	52 %	22 %	13 %	48 %
With NORS	96 %	78 %	96 %	39 %	96 %	78 %	57 %	78 %	35 %	35 %

Unlike the group of naïve NS, N2-teachers use all of the ten criteria in their impressionistic scoring. Eight of the ten criteria are used by more than 50 % of the raters, which means that there is a wide spread across the criteria for this group. Almost all raters of this group use *grammar* and *vocabulary*, while only two thirds use *pronunciation* in their explanations. 61 % focus on *fluency*, and as much as 48 % focus on *content* or factual knowledge of the candidates. 22 % use *initiative* as a criterion.

When raters use the NORS, *grammar, pronunciation and vocabulary* as well as *intelligibility* and *comprehension* stay relatively constant, while there is less focus on *fluency* and *content*. *Communicative ability* and *strategies* increase by 22 % each from impressionistic to NORS-based scoring, *initiative* sees a substantial increase in use from 22 % in the impressionistic to 78 % in the NORS-based scoring.

Table 6 Ten criteria, N2-teachers (n = 23), both scoring methods, frequencies.

Scoring method	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Impressio-nistically	3.96	1.61	3.17	1.13	2.09	1.35	0.87	0.30	0.22	0.83
With the NORS	1.04	2.17	2.74	0.52	2.57	1.39	0.91	1.43	0.48	0.57

One striking result is the N2-teachers' limited focus *pronunciation*. While the non-linguists referred to this trait approximately three times when scoring without a rating scale, N2-teachers use it only 1.61 times on average. This group actually refers to *communicative ability* (2.09) more frequently than to *pronunciation*. The most frequently used criteria for this group are *grammar* (3.96) and *vocabulary* (3.17) when scoring impressionistically. *Fluency* and *intelligibility* are used a little more than once on average, while *comprehension, initiative, strategies* and *content* are only used to a limited extent.

When basing their scores on the NORS, *grammar* declines visibly for this group as well. Now it is used less frequently than *vocabulary* (2.74), *communication* (2.57), *pronunciation* (2.17), *initiative* (1.43) and *intelligibility* (1.39). *Fluency, strategie*s and *content* are referred to less than once on average.

Table 7 Ten criteria, all SP-raters, (n = 39), both scoring methods, percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impressionis-tically	100 %	92 %	92 %	67 %	77 %	85 %	56 %	21 %	39 %	28 %
With NORS	95 %	92 %	95 %	56 %	95 %	80 %	69 %	56 %	28 %	33 %

Raters of Språkprøven (regardless of the amount of experience) refer to all ten criteria in their WR when evaluating without a rating scale. All raters of this group refer to g*rammar* and 92 % use *pronunciation* and *vocabulary*. *Intelligibility* is used by 85 % of the raters and 67 % focus on *fluency*. 39 % use *strategies* as a criterion.

When raters base their scores on the NORS, *grammar, vocabulary* and *pronunciation* stay relatively constant and continue to be used by more than 90 % of the raters. *Communicative ability* is used by almost as many raters as the formal linguistic traits: 95 % of the raters now focus on this trait when using the NORS. *Fluency* (56 %) and *strategies* (28 %) are used by fewer raters when scores are based on the NORS, while the opposite is the case for *comprehension* and *initiative* which are now used by 69 % and 56 % respectively.

Table 8 Ten criteria, all SP-raters (n=39), both scoring methods, frequencies.

Scoring method	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Impressio-nistically	4.03	2.74	3.23	1.18	2.26	1.85	1.08	0.28	0.49	0.49
With NORS	1.05	2.31	3.10	0.90	2.72	1.46	1.10	1.28	0.44	0.44

Also, when frequencies are investigated, *grammar* is the most used criterion, used more than four times on average when SP-raters score impressionistically. *Vocabulary* is used more than three times, followed by *pronunciation*, which is used almost three times on average. *Communicative ability* is also used to a considerable degree by this group, however (2.26), followed by *intelligibility* (1.85), *fluency* (1.18) and *comprehension* (1.08). *Strategies, content* and *initiative* are only used to a limited extent.

The most striking difference between scoring methods for this group, is that when using the NORS, raters only refer to *grammar* 1.05 times. *Vocabulary* (3.10) and *pronunciation* (2.31) see no

major change in use. *Fluency* is used less frequently, *initiative* more frequently when using the NORS, while the remaining traits stay relatively constant across methods.

Table 9 Ten criteria, expert SP-raters (n = 6), both scoring methods, percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impressionis-tically	100 %	100 %	100 %	67 %	83 %	83 %	83 %	0 %	67 %	17 %
With NORS	83 %	100 %	100 %	33 %	83 %	100 %	67 %	50 %	33 %	0 %

These results should be interpreted with some care as the group only consists in six raters.

Expert SP-raters are to a large extent agreed about the rating criteria of speech. When scoring impressionistically all of them focus on *grammar, vocabulary* and *pronunciation*. A large percentage focus on communicatively related criteria as well, such as *communicative ability, intelligibility, comprehension* and *strategies*: 83 % emphasise *communicative ability, intelligibility* and *comprehension*, while 67 % focus on *fluency* and *strategies*. *Initiative* is however not used by any raters of this group when scoring impressionistically.

When expert SP-raters base their scores on the NORS, the criteria *pronunciation, vocabulary* and *communication* are used by exactly the same number of raters as in the impressionistic scoring. Notably, *pronunciation* and *vocabulary* are used by 100 % of the expert SP-raters in both scoring methods. *Intelligibility* was used by 83 % of these raters when scoring impressionistically, but by all raters when using the NORS. *Fluency, comprehension, strategies* and *content* are used by less raters when using the NORS. *Initiative* was not used by any expert SP-raters when scoring impressionistically, yet while basing their scores on the NORS, 50 % of the raters of this group focus on this trait.

Table 10 Ten criteria, expert SP- raters (n = 6), both scoring methods, frequencies.

Scoring method	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Impressio-nistically	4.17	3.67	3.50	1.00	1.50	1.83	1.67	0.00	0.67	0.17
With NORS	1.17	3.17	3.17	0.67	1.83	2.67	1.50	1.17	0.33	0.00

When scoring impressionistically, the most frequently used criterion by the expert SP-raters is *grammar* (4.17). *Pronunciation* (3.67) and *vocabulary* (3.50) are the second and third most used criteria. Both *intelligibility* and *comprehension* are used a little less than two times by each rater, while

communicative ability is used even less. *Initiative,* however, is not used at all by this group when scoring impressionistically.

When using the NORS, raters of this group refer most frequently to *pronunciation* (3.17) and *vocabulary* (3.17). *Grammar* sees a dramatic decline from 4.17 to 1.17 leaving it the sixth most used criterion along with *initiative*. It is outdone by *intelligibility* (2.67), *communicative ability* (1.83) and *comprehension* (1.50). The two traits that vary the most from impressionistic to NORS-based scoring for the group of expert SP-raters, then, are *grammar, intelligibility* and *initiative.*

To facilitate a general survey of the criteria that different rater groups use, the results are presented in joint tables. Table 11 and Table 12 below present the results for all groups when scoring impressionistically, while Table 13 and Table 14 display the results for the scores based on the NORS.

Table 11 Ten criteria, all rater groups, impressionistic scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Naïve NS	100 %	100 %	92 %	75 %	42 %	75 %	58 %	0 %	0 %	8 %
N2-teachers	96 %	74 %	96 %	61 %	74 %	74 %	52 %	22 %	13 %	48 %
All SP-raters	100 %	92 %	92 %	67 %	77 %	85 %	56 %	21 %	39 %	28 %
Expert SP-raters	100 %	100 %	100 %	67 %	83 %	83 %	83 %	0 %	67 %	17 %

Table 11 shows that almost all raters of the distinct rater groups focus on *grammar*, *pronunciation* and *vocabulary* in their impressionistic scoring. N2-teachers stand out from the other groups in that only 74 % focus on *pronunciation*, which is more than 20 % less than the other groups. I*ntelligibility* is also used by a large percentage of the informants when scoring without a rating scale. Raters of Språkprøven (all SP-raters as well as expert SP-raters) use this criterion to a somewhat higher degree than informants without rater training (naïve NS and N2-teachers). *Fluency* is used a little more by the naïve NS than by raters of the other groups, but the differences between groups are minor. Some traits are however used quite differently across rater groups. *Communicative ability*, for example, is used by 83 % of the expert SP-raters in their impressionistic scoring, as opposed to 42 % of the naïve NS. Similarly, 83 % of the expert SP- raters focus on *comprehension*, as compared to only approximately half of the raters of the other groups. *Initiative* is not used by any of the naïve NS nor by the expert SP-raters, yet by slightly more than 20 % of the N2-teachers and all SP-raters. The reference to *strategies* is also quite interesting when seen

across rater groups: for this trait there is a clear progression from 0 % of the naïve NS, through 13 % of the L2-teacher group, 39 % of all SP-raters and finally 67 % of the expert SP-raters. The reference to the criterion *content* also varies a great deal across rater groups: it is used by almost half of the raters of the L2-teacher group and by almost one third of all the raters of Språkprøven. Only 17 % of the expert SP-raters refer to it, and as few as 8 % of the naïve NS.

Table 12 Ten criteria, all rater groups, impressionistic scoring, frequencies.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Naïve NS	4.08	3.08	3.83	1.33	1.00	1.75	1.42	0.00	0.08	0.08
N2-teachers	3.96	1.61	3.17	1.13	2.09	1.35	0.87	0.30	0.22	0.83
All SP-raters	4.03	2.74	3.23	1.18	2.26	1.85	1.08	0.28	0.49	0.49
Expert SP-raters	4.17	3.67	3.50	1.00	1.50	1.83	1.67	0.00	0.67	0.17

Comment to the naïve NS results for strategies and content:

When comparing the degree to which different rater groups use the criteria, we see that the expert SP-raters use four of the criteria more frequently than the other groups. This is the case for *grammar, pronunciation, comprehension* and *strategies*. *Fluency* and *initiative*, on the other hand, are less used by this group than by the other groups.

There are only minor differences between the groups regarding the criterion *grammar*. It is, in fact, the most frequently used criterion across rater groups. Except for the N2-teachers, raters of all groups refer to this trait more than four times when scoring impressionistically. *Pronunciation* is used more or less three times on average by all raters, except for the N2-teachers who only refer to it 1.61 times. *Vocabulary* does not vary much across groups either (from 3.17 to 3.83). Nor do *fluency* and *intelligibility*. The expert SP-raters as a group use *fluency* less than the other groups. *Communicative ability* is referred to 1.50 times on average by the expert SP-raters, which is less than the N2-teachers (2.09), as well as being less than all SP-raters (2.26). *Comprehension*, on the other hand, is most frequently used by the group of expert SP-raters: N2-teachers (0.87) refer to *comprehension* only half as often as the expert SP-raters (1.67). Naïve NS refer to it 1.42 times and all SP-raters 1.08 times. *Initiative* is not used at all by the expert SP-raters or by the group of naïve NS. *Strategies* ranges from zero to 0.67 from the naive to the expert SP-raters, which is parallel to the pattern of the percentage calculation.

Table 13 Ten criteria, all rater groups, NORS-based scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Naïve NS	100 %	92 %	100 %	42 %	58 %	75 %	67 %	58 %	8 %	0 %
N2-teachers	96 %	78 %	96 %	39 %	96 %	78 %	57 %	78 %	35 %	35 %
All SP-raters	95 %	92 %	95 %	56 %	95 %	80 %	69 %	56 %	28 %	33 %
Expert SP-raters	83 %	100 %	100 %	33 %	83 %	100 %	67 %	50 %	33 %	0 %

When using the NORS, *grammar* is used by a lower percentage of expert SP-raters than by raters of the other groups: only 83 % of the expert SP-raters focus on grammar, while 100 % of the naïve NS and approximately 95 % of the N2-teachers and all SP-raters do the same. *Pronunciation* is used by all the expert SP-raters and by more than 90 % of the naïve NS and all SP-raters, while only 78 % of the N2-teachers focus on this. *Vocabulary* is the criterion used by the largest number of raters when scores are based on the NORS: all raters of the groups of the naïve NS and the expert SP-raters refer to it, 96 % of the N2-teachers and 95 % of the group of all SP-raters. *Fluency* varies between 33 % (expert SP-raters) and 56 % (all SP-raters) and *communicative ability* is used by a large percentage of raters of all groups (from 83 % to 96 %), the exception being the naïve NS group where only 58 % use the criterion. *Intelligibility* sees an increase according to the level of experience of raters: 75 % of the naïve NS refer to it, 78 % of the N2-teachers, 80 % of all SP-raters and finally 100 % of the expert SP-raters. When basing the scores on the NORS, *comprehension* and *initiative* are then used by more than 50 % of the raters regardless of background. For *comprehension,* the group of N2-teachers stands out from the rest: naïve NS, SP-raters and expert SP-rater all obtained a percentage above 67 %, while in the L2-teacher group only 57 % refer to this criterion.

N2-teachers differ from the rest when it comes to the criterion *initiative* as well. The number of N2-teachers referring to initiative is higher than the other groups: 78 % of the N2-teachers use this trait, while the percentages for the other groups vary between 50 % and 58 %. Strategies are used by approximately 30 % of the raters of N2-teachers, all SP-raters and expert SP-raters. Here the group of naïve NS differs from the rest: only 8 % of these raters refer to strategies in their argumentation. None of the raters of the groups of naïve NS and expert SP-raters uses *content* as a criterion. For comparison, in the groups of N2-teachers and all SP-raters, the percentage is above 30 %.

Table 14 Ten criteria, all rater groups, NORS-based scoring, frequencies.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Naïve NS	1.00	2.92	4.00	0.75	1.50	2.42	1.42	1.83	0.08	0.08
N2-teachers	1.04	2.17	2.74	0.52	2.57	1.39	0.91	1.43	0.48	0.57
All SP-raters	1.05	2.31	3.10	0.90	2.72	1.46	1.10	1.28	0.44	0.44
Expert SP-raters	1.17	3.17	3.17	0.67	1.83	2.67	1.50	1.17	0.33	0.00

Also when using the NORS, the expert SP-raters use several of the criteria to a larger extent than the other groups: this is the case for *grammar*, *pronunciation* and *comprehension* and *intelligibility*. Naïve NS top the frequency list for the traits *vocabulary* and *initiative*, while N2-teachers use *strategies* and *content* more than the other groups when scores are based on the NORS. The group of all SP-raters uses *fluency* and *communicative ability* to a greater extent than the other groups.

Grammar and *fluency* are used by a similar number of raters across groups. *Pronunciation* varies between 2.17 and 3.17, N2-teachers referring to it the least and expert SP-raters the most. *Vocabulary* is used approximately three times on average by all groups, except for the naïve NS who use it four times on average. Regarding *communicative ability* there are variations in use between 1.50 for the naïve NS and 2.72 for the group of all SP-raters. N2-teachers refer to this trait 2.57 times while the expert SP-raters only refer to it 1.83 times. When it comes to the use of *intelligibility,* on the other hand, the expert SP-raters use it most frequently: They refer to *intelligibility* 2.67 times, followed by the group of naïve NS (2.42), all SP-raters (1.46) and finally N2-teachers (1.39). *Comprehension* varies between 0.91 and 1.50 times, N2-teachers on the bottom and expert SP-raters on the top of the list. For *initiative*, on the other side, the expert SP-raters use the criterion less than the other groups, but the differences are rather small: the naïve NS refer to *initiative* 1.83 times on average, while the expert SP-raters use it 1.17 times. *Strategies* are used between 0.33 and 0.48 times on average by raters of all groups with the exception of the naïve NS group, which only refers to it 0.08 times on average. *Content* is not used by the expert and the naïve NS groups at all, the remaining groups refer to it less than once on average.

The results so far indicate that there are some differences between the groups regarding the criteria raters use in their evaluations of L2-speech. There seems to be a greater degree of agreement between raters for the formal linguistic traits, *grammar, pronunciation* and *vocabulary*, than for the aspects regarding communicative functionality, such as *communicative ability, intelligibility,*

comprehension, initiative and *strategies*. Whether this is the case, may be investigated by merging formal linguistic and communicatively related traits into two separate categories as presented in the tables below. *Formal linguistic traits* are here defined as those aspects of speech that relate to a norm of correctness. *Grammar* is the most obvious of these traits, as there is a written norm for grammatically correct use of Norwegian (Hagen 1998, Faarlund et al. 1997). *Pronunciation* lacks a written norm for correct use, and the dialectal variation in Norway is extensive. Yet, it is reasonable to assume that NS do have a feeling for which pronunciation is within the limits of acceptable variation for Norwegian and which is not. *Vocabulary* is also classified as a formal linguistic trait. As for grammar and pronunciation it is possible to relate the use of words to a norm of correctness: it makes sense to talk about a correct choice of words in a context, at least to some degree.

As opposed to formal linguistic traits are the traits classified under the label *communicative functionality*. These traits do not relate to a norm of correctness, rather it is a question of whether or not something is functional or effective in a communication situation, that is whether the candidate manages to get the meaning across, despite formal errors. The traits grouped in this category are *communicative ability, intelligibility, comprehension, initiative* and *strategies*. Even though there is not an opposition between formal correctness and communicative functionality, in the sense that a formally correct language will probably always be more functional than an erroneous language, it is useful to draw this distinction when investigating raters' evaluations. *Fluency* and *content* are left out: It is hard to tell whether raters use fluency by reference to formal correctness or communicative functionality. Content is left out because it is not considered a linguistic skill at all: Rather it should be classified as a person's knowledge of the world (Bachman 1990) or topical knowledge (Bachman and Palmer 1996).

Table 15 Formal linguistic traits, all rater groups, both scoring methods, percentages.

Average of percentages	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Impressionistic scoring	97 %	89 %	95 %	100 %
NORS-based scoring	97 %	90 %	94 %	94 %

*(The percentages are calculated by adding up the number of raters of each rater group focusing on grammar, pronunciation and vocabulary (*Table 11 *and* Table 13*) and dividing it by three (criteria). This gives the average percentage of raters focusing on formal linguistic traits).*

As can be seen in Table 15, there are only minor differences between the groups for the use of formal linguistic traits. The percentage of raters referring to such traits is above 90 % for all

groups for both scoring methods, with the exception of the group of N2-teachers when scoring impressionistically. This reflects N2-teachers' lack of focus on *pronunciation* as discussed earlier in this chapter.

Table 16 Formal linguistic traits, all rater groups, both scoring methods, frequencies.

Frequencies of use	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Impressionistic scoring	3.66	2.91	3.33	3.38
NORS-based scoring	2.64	1.97	2.15	2.50

(The numbers are calculated by adding up the frequencies of grammar, pronunciation and vocabulary displayed in Table 13 and Table 14 divided by three (criteria) to get the average of frequency by which formal traits are used).

The frequency table shows similar results. There are no major differences between the groups, except between that of N2-teachers and the others as mentioned above. There is however a noticeable decline for all rater groups in how often raters refer to formal traits when using the NORS.

Let us now turn to the tables presenting the results of focus on communicative functionality.

Table 17 Communicative functionality, all rater groups, both scoring methods, percentages.

Averages of percentages	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Impressionistic scoring	35 %	47 %	56 %	63 %
NORS-based scoring	53 %	69 %	66 %	67 %

As Table 17 shows, there are fairly substantial differences between the groups regarding the percentages of raters focusing on communicative functionality. The differences are particularly large for the impressionistically based scores where there is a clear progression from the group of naïve NS to the group of expert SP-raters: in the group of naïve NS only an average of 35 % of the raters focus on communicatively related criteria. For the expert rater group, on the other hand, 63 % or almost twice as many raters, focus on these aspects. In the group of all SP-raters, 56 % focus on communicative functionality, while 47 % of the N2-teachers do the same. When basing their scores on the NORS these group differences are almost eliminated. Now, there are only minor differences between the groups of N2-teachers, all SP-raters and experienced SP-raters. The naïve NS stand out from the rest, though, since only 53 % of the raters of this group

refer to communicative functionality in their argumentation, while in the other groups the average percentage is between 66 % and 69 %. Table 18 below displays the frequencies by which raters of different groups refer to communicative functionality.

Table 18 Communicative functionality, all rater groups, both scoring methods, frequencies

Total of frequencies	Naïve NS	N2-teachers	Raters of Språkprøven	Expert raters of Språkprøven
Impressionistic scoring	4.17	4.31	5.19	5.00
NORS-based scoring	5.71	6.78	7.00	7.50

The pattern apparent in Table 17 is similar when frequency of use is considered. Even though the differences between the groups are not as large as when percentages were under the spotlight, different rater groups refer to communicatively related criteria to a somewhat different degree. Again, there is a cline in frequency from the naïve NS to the expert SP-raters, naïve NS referring least frequently and expert SP-raters most frequently to communicative aspects of speech. This pattern is particularly striking when raters base their scores on the NORS: the naïve NS refer to such traits 5.71 times on average, while the expert SP-raters refer to them 7.50 times. When scoring impressionistically the group of all SP-raters tops the list of most frequent reference to communicative effect (5.19) followed by the group of expert SP-raters (5.00). N2-teachers and naïve NS refer to it 4.31 and 4.17 times respectively.

If summing up the results of the study of group differences and the use of criteria, we see that they do to some degree support the assumption that different rater groups focus on different criteria in their WR. These differences are of three kinds: firstly, there are differences in the number of criteria, or the spread of criteria, used by different rater groups. Secondly, there are differences in the internal agreement of different rater groups regarding the criteria for L2-speech. And thirdly, there are differences in the specific criteria different rater groups utilise, i.e. various rater groups tend to focus on somewhat different aspects of speech.

10.1.1. Number of criteria used by distinct rater groups

When scoring impressionistically, the group of expert raters of Språkprøven uses a wider range of criteria than the other groups. This is confirmed when percentages of raters who use the traits, as well as when frequencies of reference to a trait are considered. As seen in Table 11 four of the ten criteria are used by a larger percentage of experienced raters than by other rater groups, and

for two more criteria expert SP-raters top the list together with one other group. When frequencies are considered the pattern is less clear, but still this group uses four of the criteria more frequently than the other rater groups. When using the NORS the table displaying the percentage of raters using the different criteria, does not show any major differences between the number of raters using distinct criteria (Table 13). The expert SP-rater group does not differ from the other groups as to the number of criteria used. When mean frequencies are considered, on the other hand (Table 14), the group of expert SP-raters uses four of the ten criteria more frequently than any other group. No other group tops the list for that many criteria. Hence, there seems to be a larger spread across criteria amongst the expert SP-raters than amongst raters of any other group.

10.1.2. Internal agreement of criteria

Despite the use of a wider range of criteria and more frequent reference to the criteria amongst expert SP-raters than amongst raters of the other groups, expert SP-raters are relatively internally consistent about the criteria they use. There is full agreement (100 % or 0 % of the raters) for as many as four of the ten criteria when scoring impressionistically, as well as when scores are based on the NORS. This is not very surprising, though: one would assume rater training and extensive rater experience to make raters in accordance with one another, not only with reference to the scores they give but to the traits they use as well. In Section 10.3 we will investigate the match between the criteria used by the different groups and the criteria of in the rating scale of Språkprøven. For now, suffice it to state that the group of the most experienced raters shows higher internal consistency of criteria than the other rater groups.

More surprisingly, raters of the group of naïve NS are also highly agreed about the criteria for judging speech. When scoring impressionistically, four criteria are referred to by all or no raters of this group (Table 11). When they base their scores on the NORS, three criteria are used in the same way by all naïve NS. For comparison, in the groups of N2-teachers and the group of all SP-raters no criteria are used by 100 % or 0 % of the raters regardless of scoring method, the only exception being all SP-raters' focus on grammar when scoring impressionistically. This is quite interesting given the fact that the naïve NS have different educational backgrounds, no experience as N2-teachers or testers, and they have no experience whatsoever with Språkprøven or the NORS. Yet, they seem to have a similar intuition about the components of L2-speech. This will be discussed in more detail in Chapter 11.

10.1.3. The focus on formal linguistic traits versus communicative functionality

The results seem to indicate that rater training and experience affect the criteria that raters use, especially when the communicatively related aspects of speech are considered. When scoring impressionistically, all groups focus on the formal linguistic traits to a large extent, the exception being N2-teachers' relative lack of focus on *pronunciation*. When it comes to communicatively related criteria, such as *communicative ability, intelligibility, comprehension, initiative* and *strategies*, on the other hand, there is a substantial degree of variation between groups. In particular, there are differences between the groups of naïve NS and expert SP-raters: a large percentage of raters of the expert group focus on communicatively related criteria, while the naïve NS focus almost exclusively on formal aspects of performance.

Even though the comparison of rater groups showed several similarities between the naïve NS and the expert-rater groups for internal agreement of criteria, Table 17 and Table 18 show that there are however some important differences between these groups when it comes to the criteria they use in their evaluations. Naïve NS are more exclusively focused on formal linguistic traits, while the groups of experienced raters also base their scores on the degree to which the candidates manage to cope with the communicative task with which they are faced.

10.2. Do raters focus on different criteria when they score impressionistically and NORS-based?

In this section I investigate the effect of rating method on the criteria raters use in their evaluations. Hence, the main focus is no longer on differences between groups but rather on differences in the use of criteria depending on whether or not raters base their scores on the NORS or not. Of course, this question cannot be investigated entirely independently of rater groups, so there is a danger of some repetition from Section 10.1.

Table 19 Difference between scoring methods, ten criteria, all raters (n=74), percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impressionis-tically	99 %	88 %	93 %	66 %	70 %	80 %	55 %	18 %	24 %	31 %
With NORS	96 %	88 %	96 %	49 %	89 %	78 %	65 %	64 %	27 %	28 %
Difference	- 3%	-	+ 3 %	- 17 %	+ 19 %	- 2 %	+ 10 %	+ 46 %	+ 3 %	- 3 %

The last row refers to the difference in percentages of raters using a certain criterion from impressionistic to NORS-based scoring. Low numbers indicate that the criterion stays almost invariable/ across scoring methods while high numbers indicate the opposite. Plus and minus signs indicate whether there has been a decline (- sign) or an increase (+ sign) from impressionistic to scale–based scoring.

When scoring impressionistically, close to 100 % of the raters focus on the formal linguistic traits: *grammar* (99 %) and *vocabulary* (93 %). *Pronunciation* (88%) is used a little less than the other two formal traits, which is probably due to the lesser focus on pronunciation by the group of N2-teachers as discussed earlier. 80 % of the raters focus on *intelligibility*, while 70 % emphasise the *communicative abilities* of candidates. More than half of the raters (66 %) use *fluency* as a criterion when they score without an explicit rating scale at hand, while less than one third of the raters employ *strategies* (24 %) and *content* (31 %) in their WR. The criterion which is least used by all the raters is *initiative*. Only 18 % of the raters refer to this aspect when scoring impressionistically.

When raters use the NORS as a guide, some of the criteria see quite a considerable change in use while others stay more or less constant. The formal linguistic traits continue to be used by almost all raters. Likewise, there are only minor differences in use for the criteria *intelligibility* (-2), *strategies* (+3) and *content* (-3). The four remaining traits, however, are used by quite a different amount of raters depending on rating methods: this is the case for *comprehension, fluency, communicative ability* and *initiative*. As already mentioned, *fluency* was used by 66 % of the raters when scoring impressionistically. When using the NORS, the amount of raters focusing on this trait drops with by 17 %. The other traits that are used differently are categorised as communicatively related criteria. These criteria are all put to use to a larger extent when raters use the NORS than when scoring impressionistically. *Comprehension* increases from 55 % to 65 %, and *communicative ability* sees an increase from 70 % to 89 %. That makes communicative ability the third most frequently used criterion after grammar and vocabulary when scores are based on the NORS. The most astonishing difference in use, though, is found for the criterion *initiative*: only 18 % of the raters referred to this criterion when scoring impressionistically, while as many as 64 % used it when basing their scores on the NORS. This gives an amazing increase of 46 %!

Table 20 Difference between scoring methods, ten criteria, all raters (n=74), frequencies.

Scoring method	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strategies	Content
Impressio-nistically	4.01	2.45	3.31	1.19	2.00	1.68	1.07	0.24	0.34	0.53
With NORS	1.04	2.36	3.14	0.76	2.47	1.59	1.09	1.42	0.39	0.41
Difference	- 2.97	- 0.09	- 0.17	- 0.43	+ 0.47	- 0.09	+ 0.02	+ 1.18	+ 0.05	- 0.12

Grammar is the criterion used most frequently by all raters when scoring impressionistically. Each rater refers to this aspect on average 4.01 times. *Vocabulary* (3.31) and *pronunciation* (2.45) are the

second and third most used criteria, followed by *communicative ability* (2.00) and *intelligibility* (1.68). *Fluency* and *comprehension* are used about once by each rater, while *initiative, strategies* and *content* have very low frequencies. As in Table 19, *initiative* (0.24) is the least used criterion of all.

As was the case when percentages were considered, the picture changes when raters use the NORS. While *pronunciation* and *vocabulary* stay quite constant as two of the most used criteria, *grammar* sees a dramatic decline in use: from being used approximately four times by each rater, *grammar* now drops to a frequency of 1.04. From being the most frequently used criterion when raters score on impressionistic grounds, *grammar* is only the seventh most used criterion when the NORS is used as a guide. Quite the opposite is the case for *initiative*: from being the least used criterion, it increases to the fifth most used trait with a frequency of 1.42. In other words, when the NORS is the basis for decisions, *initiative* is used more often than *grammar*. Not surprisingly, the use of the NORS also leads to an increased focus in use of the criterion *communicative ability*: when scores are based on the rating scale, *communicative ability* is the second most used criterion after *vocabulary*. The criteria that vary the most from impressionistic to NORS-based evaluation, then, are *grammar* (- 2.97) and *initiative* (+ 1.18). *Fluency* (- 0.43) and *communicative ability* (+ 0.47) vary to some extent, while the rest of the criteria are used with approximately the same frequencies across scoring methods.

It is interesting to draw a line between formal traits and communicative functionality in relation to the effect of scoring method on the use of criteria, as well.

Table 21 Formal linguistic traits, all raters (n = 74), both scoring methods, percentages and frequencies.

Formal linguistic traits.		
All raters	Average of percentages	Frequency of use
Impressionistic scoring	95 %	3.32
NORS-based scoring	94 %	2.32

Based on the numbers of Table 15 and Table 16.

When raters score impressionistically, 95 % of all raters focus on the formal traits, *grammar, pronunciation* and *vocabulary*. They do so on average 3.32 times. When the NORS is used as a basis, the percentage of raters focusing on these traits stays close to constant, but each rater however refers to the traits less frequently.

Table 22 Communicative functionality, all raters (n=74), both scoring methods, percentages and frequencies.

Communicative functionality.		
All raters	Average of percentages	Frequency of use
Impressionistic scoring	50 %	4.67
NORS-based scoring	64 %	6.75

The communicatively related traits are used by 50 % of all raters when scoring without a rating scale. Raters who use these traits, do so quite frequently, though: communicative functionality is referred to on average 4.67 times. When the NORS is applied, the percentages of raters referring to the trait, as well as the frequency by which they do so, rise considerably: 64 % of the raters make use of these traits in their NORS-based evaluation, and they are used to a great extent: each rater refers to these traits on average 6.75 times.

So far in this chapter, the variables *rater training* and *scoring method* have been treated separately. Yet, the variables are highly inter-related, and any drawing of a clear line between them will necessarily be a simplification. It would therefore be interesting to look at the effect of rater training and of rating scale on the criteria in combination. The question whether some rater groups vary more from impressionistic to NORS-based ratings than other groups will be investigated. Without formulating any specific hypothesis, it is inevitable to have some expectations as to the results. The expected results would be for the naïve NS-group to vary the most and the expert group the least from impressionistic to NORS-based scorings. This would be a logical consequence of trained and experienced raters having internalised the rating criteria.

Table 23 Difference between scoring methods, ten criteria, all rater groups, percentages, total variance.

Rater groups	Gram-mar	Pronun-ciation	Voca-bulary	Fluency	Commu-nicative ability	Intel-ligibi-lity	Compre-hension	Initia-tive	Strategies	Content	**Total variance for rater groups**
Naïve NS	-	-8 %	+8 %	- 33 %	+ 16 %	-	+ 9 %	+ 58 %	+ 8 %	- 8 %	**148**
N2-teachers	-	+4 %	-	- 22 %	+ 22 %	+2 %	+ 5 %	+ 56 %	+ 22 %	- 13 %	**146**
All SP-raters	-5 %	-	+3 %	- 11 %	+ 18 %	-5 %	+ 13 %	+ 35 %	- 11 %	+ 5 %	**106**
Expert SP-raters	-17 %	-	-	- 34 %	-	+17 %	- 16 %	+ 50 %	- 34 %	- 17 %	**185**
											585

Looking at the numbers of Table 23, some patterns come to sight: for the trait *grammar,* the two groups of raters of Språkprøven differ from the groups of non-raters: the group of expert SP-raters declines by 17 % from impressionistic to scale–based scorings, the group of all SP-raters with 5 %. The naïve NS and N2-teachers stay constant across rating methods for this trait. *Pronunciation* and *vocabulary* see only minor differences between groups across methods. *Fluency,* on the other hand, drops considerably from impressionistic to NORS-based scorings for all groups. The distinct groups do however vary to a different extent for this trait. Naïve NS and expert-raters both see a dramatic decline of more than 30 %. N2-teachers only decrease by 22 % while the group of all raters drops by 11 %. (Given the fact that the two groups of SP-raters are not discrete, these 11 % overlap with the results of the expert group). Interestingly, *communicative ability*, sees an increase for all groups with the exception of the group of the expert SP-raters. In this group a large percentage of raters already focused on this trait when scoring impressionistically, as opposed to the other groups (see Table 9). *Intelligibility* stays relatively constant for all groups except for the group of expert SP-raters, which sees an increase of 17 %. For the trait *comprehension* the expert SP-raters once more stand out from the rest. While in the other groups the number of raters focusing on this trait rises (from 5 % to 13 %), the group of experienced raters sees a decline in use of 16 %. This may be caused by the relatively large number of expert SP-raters focusing on this trait in their impressionistic scoring (83 %, Table 11) and the fact that it is not explicitly mentioned in the rating scale. *Initiative* shoots up by more than 50 % for all rater groups, the only exception being the group of all SP-raters, which has an increase of only 35 % mirroring the salient position of this trait in the NORS. *Strategies* are used rather differently by the various rater groups in as much as that while the groups of non-SP-raters use this trait more when basing their scores on the NORS, the opposite is the case for the two groups of SP-raters. Expert SP-raters differ the most from impressionistic to NORS-based scoring with a decrease of 34 % for this trait. Finally, *content,* varies from – 8 % (naïve NS) to –17 % (expert SP-raters). All SP-raters, on the other hand, see a small rise in use of this criterion.

The column to the very right of the table displays the total variance of each rater group across rating criteria. The results very clearly undermine the expected results namely that the group of expert SP-raters does not vary the least form impressionistic to NORS-based scoring, as assumed if the rating criteria of the NORS had been internalised. In fact, the expert group varies more than any of the other groups (185 points). Naïve NS and N2-teachers vary to a similar degree with 148 and 146 points respectively, while the group of all SP-raters varies the least with 106

points. These results of variability should be interpreted with some caution, though. Some rater groups may be more explicit in their WR than others, that is, they may actually refer to more traits than other groups. This should not, however, affect the degree of difference between the two scoring methods: if raters of one particular group tend to give more abundant reports of the scores they give, there is no reason to assume that this would only be the case for one scoring method and not for the other. Therefore when differences between the two scoring methods are considered for various rating groups, this should not, as far as I am aware, have any major relevance the results.

Table 24 Difference between scoring methods, ten criteria, all rater groups, frequencies, total variance.

Rater groups	Gram-mar	Pronu n-ciation	Voca-bulary	Fluency	Commu-nicative ability	Intelligi-bility	Compre-hension	Initiative	Strate-gies	Content	**Total variance for rater groups**
Naïve NS	-3.08	-0.16	+0.17	-0.58	+0.50	+0.67	-	+1.83	-	-	**6.99**
N2-teachers	-2.92	+0.56	-0.43	-0.61	+0.48	+0.04	+0.04	+1.13	+0.26	-0.26	**6.73**
All SP-raters	-2.98	-0.43	-0.13	-0.28	+0.46	-0.39	+0.02	+1.00	-0.05	-0.05	**5.79**
Expert SP-raters	-3.00	-0.50	-0.33	-0.33	+0.33	+0.84	-0.17	+1.17	-0.34	-0.17	**7.18**

The two most striking patterns apparent in Table 24 are those regarding grammar on the one hand and initiative, on the other. As Table 20 showed, g*rammar* is the trait that varies the most from impressionistic to NORS-based scoring. Nevertheless, there are no major group differences for this trait: all groups drop by about 3.00. In addition to *grammar*, *initiative* is the trait that is used most differently depending on scoring methods. It sees a considerable increase for all rater groups: the naïve NS have the greatest rise (+ 1.83) which is understandable taking into consideration the fact that they did not refer to this criterion at all in their impressionistic scoring. *Pronunciation* also decreases in use for all groups with the exception of N2-teachers who, on the contrary, use this trait a little more often when using the rating scale. This may be due to the relatively infrequent use of pronunciation for this group when scoring impressionistically. *Vocabulary* declines for all groups (from – 0.13 to – 0.43) with the exception of the naïve NS who show a small increase (+ 0.17) in use for this trait. *Fluency* sees a reduction in use for all rater groups, a bit more for the groups of naïve NS and N2-teachers than the two groups of SP-raters. *Communicative ability* changes in the opposite direction to *fluency:* raters of all groups tend to refer to this trait more often when scores are based on the NORS. The group of expert SP-raters also uses this trait to a larger extent now, but the group varies less than the other rater groups. The

use of *intelligibility* increases the most by the groups of expert SP-raters (+ 0.84) and naïve NS(+ 0.67). N2-teachers use the trait to a similar extent across scoring methods while the group of all SP-raters shows a decrease of 0.39. *Comprehension* stays rather constant from the first scoring method to the next, and there are only minor differences between groups. *Strategies* and *content* are little used criteria in both scorings, and the differences between the groups are small for these traits.

Table 25 Difference between scoring methods, formal linguistic traits, all rater groups, percentages.

	Formal linguistic traits.			
	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Difference in percentages from impress. to NORS-based scoring	-	+ 1 %	- 1 %	- 6 %

Based on the numbers of Table 15.

As evident in the above table, there are only minor differences between the number of raters focusing on *formal traits* from impressionistic to NORS-based scoring. In the group of naïve NS there is no change at all, while N2-teachers and all raters of Språkprøven vary with one percent each in the opposite direction. Expert raters of Språkprøven, however, vary the most across scoring methods: 6 % fewer raters focus on formal linguistic traits when scorings are based on the NORS than when scoring impressionistically.

Table 26 Difference between scoring methods, formal linguistic traits, all rater groups, frequencies.

	Formal linguistic traits.			
	Naïve NS	N2-teachers	Raters of Språkprøven	Expert raters of Språkprøven
Difference in frequency from impress. to NORS-based scoring	- 1.02	- 0.94	- 1.18	- 0.88

Based on the numbers of Table 16.

There are larger differences in frequency of use than in the amount of raters using the formal traits between the two scoring methods. In other words, even though the number of raters

focusing on formal traits stays almost constant across scoring methods, raters use these criteria less frequently in their WR when using the NORS. The group that varies the most for these traits is the group of all SP-raters followed by the group of naïve NS. N2-teachers and expert raters of Språkprøven show a more modest degree of variation with 0.94 and 0.88 respectively. Expert SP-raters therefore seem to be the most stable group for formal traits where frequency of use is concerned.

Table 27 Differences between scoring methods, communicative functionality, all rater groups, percentages.

	Communicative functionality.			
	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Difference in percentages from impress. to NORS-based scoring	+ 18 %	+ 22 %	+ 10 %	+ 4 %

Based on the numbers of Table 17.

The results shown in Table 27 are very interesting indeed! While there were only minor differences between scoring methods for percentages of raters focusing on formal traits, the case is quite different for communicative functionality. All groups use communicative functionality to a larger extent when scoring with the NORS than when scoring impressionistically, and for some groups the differences are quite striking. For example, for the group of N2-teachers there is a sharp increase of 22 %. Similarly, the naïve NS see a rise of 18 %. The two groups of raters of Språkprøven also vary from one scoring method to the other but the differences are much smaller than for the groups of untrained raters. All raters of Språkprøven increase by 10 %, but the most experienced raters by only 4 %. Hence, as opposed to the case for formal linguistic traits, the group of expert SP-raters is now the most constant one.

Table 28 Difference between scoring methods, communicative functionality, all rater groups, frequencies.

Communicative functionality.				
	Naïve NS	N2-teachers	All raters of Språkprøven	Expert raters of Språkprøven
Difference in frequency from impress. to NORS-based scoring	+ 1.54	+ 2.47	+ 1.81	+ 2.50

Based on the numbers of Table 18.

When frequency of use is considered there is also a quite substantial difference between scoring methods for communicative functionality. All groups refer more often to this aspect when using the NORS than when scoring impressionistically. Expert SP-raters show the steepest increase (+ 2.50), followed by N2-teachers, (+ 2.47). All raters of Språkprøven vary by + 1.81 while the group of naïve NS refers 1.54 times more to this in their NORS-based scorings. So, even though the group of expert SP-raters and all raters of Språkprøven are more focused on the communicative aspects of speech in their impressionistic scoring, they also increase the most from impressionistic to NORS-based scoring when frequency of use is concerned. This will be further treated in the discussion of Chapter 11.

Summing up the main differences between scoring methods some principal features may be distinguished. Firstly, where percentages of raters are concerned, the formal traits *grammar, pronunciation* and *vocabulary* stay constant across scoring methods for all groups. *Grammar* is however used less frequently by all groups when scores are based on the NORS than when raters evaluate impressionistically. The trait that varies the most from impressionistic to NORS-based evaluation is *initiative:* It increases in use by 46 % of raters and it is used 1.18 times more frequently when raters base their scores on the NORS. Communicative functionality is used to a much larger extent when raters use the NORS than when scoring impressionistically: the percentages of raters using the communicatively related criteria rise from 50 % to 64 % from impressionistic to NORS-based scoring, and the frequency of use rises from 4.67 to 6.75 when all raters are seen together. Where percentages of raters are concerned, the groups that vary the most between scoring methods for communicative functionality are the two groups of inexperienced raters, i.e. the naïve NS and the N2-teachers. The expert SP-raters do however vary a great deal when it comes to the frequency by which such aspects are used.

10.3 The effect of rater training and rating scale on construct validity.

Finally, we have reached the section in which hypotheses H3 and H4 are tested against the qualitative data. As mentioned in the introduction of this chapter, the main purpose of the qualitative study has been to investigate whether rater training and the use of an explicit rating scale affect construct validity positively. The match between raters' criteria and the criteria of the NORS is used as an index of construct validity. The expected results would be that both rater training and the use of the NORS have a positive effect on construct validity as presented graphically in Figure 11. This implies that the greatest indexes of validity should be found in the group of expert SP-raters when scores are based on the NORS, while the group of naïve NS (and N2-teachers) when scoring impressionistically would show the poorest indexes of validity.

10.3.1 Testing of H3: The effect of rater training on construct validity.

Hypothesis 3 regards the effect of rater training on construct validity of test scores and states that:

- H3: **Training of raters** affects **construct validity** (defined as the match between the criteria of the scale and those of the raters) positively: there is a greater match between the criteria of the NORS and those of the trained raters than between the NORS and the criteria used by other rater groups.

H3 is tested against the empirical data presented in Table 11,Table 12, Table 13, Table 14, Table 15,Table 16, Table 17 and Table 18.

According to the NORS, raters' main focus should be on the *communicative ability* and *initiative* of the candidates. Raters should also pay attention to the formal traits: *grammar, pronunciation* and *vocabulary,* but connected closely with whether or not errors hinder communication. *Intelligibility* and *comprehension* are only implicit in the scale descriptors. *Strategies* are also only implicitly mentioned, while *fluency* and *content* are not mentioned in the NORS at all. A valid evaluation based on this scale would therefore be one in which the ability to get the meaning across in an understandable way is more important than formal correctness.

The section is organised in the following way: Firstly, I look at the use of the criteria explicit in the NORS (*communicative ability* and *initiative, grammar, vocabulary* and *pronunciation*) and the extent to

which raters of different groups use these traits in their WR. After having looked at these traits separately, formal linguistic traits and communicatively related traits are grouped as in Section 10.2 above. Then, the two traits used by many raters despite the fact that they are not included in the NORS, namely *fluency* and *content*, are studied in relation to the difference in use from one scoring method to the other.

Again, when focus is on the effect of rater training on test scores, group differences are under study. As shown in Table 11 close to 100 % of raters of all rater groups focus on the formal linguistic traits in their impressionistic scoring. There are no major differences between the groups for these traits, except for the N2-teachers who focus less on pronunciation than the other groups. Formal linguistic traits are used more frequently than any other trait by all rater groups (again with the exception of the N2-teachers' relatively infrequent use of *pronunciation).*

Communicative ability is used by 83 % of the expert SP-raters, by more than 70 % of the group of all SP-raters and by N2-teachers, but, strikingly enough, only by 42 % of the raters of the naïve NS- group. The frequency table (Table 12) shows that this groups also refers less frequently to this trait, followed actually by the group of expert SP-raters, while the group of all SP-raters uses it most frequently. When basing their scores on the NORS, the percentages rise for all groups for this trait (Table 13) but still, in the group of naïve NS, the amount of raters focusing on this trait is about half the amount in the other groups. The trait is most frequently used by the group of all SP-raters and by N2-teachers, followed by the group of expert SP-raters. Again the group of naïve NS refers less frequently to this trait. For communicative ability, then, there seems to be a positive correlation between rater experience and the use of the criteria of the NORS. This is in favour of H3.

Initiative is the other principal criterion highlighted in the NORS. For a support of H3, raters of Språkprøven should focus more on this trait than informants without rater training (naïve NS and N2-teachers). When scoring impressionistically, *initiative* is only used to a very limited degree by all rater groups. No rater groups refer to this trait more than 0.30 times. In the group of naïve NS no raters refer to this trait, compared to a little more than 20 % of the N2-teachers and all SP-raters. Surprisingly enough, 0 % of the expert SP-raters refer to this trait when scoring impressionistically. When using the NORS, on the other hand, the use of *initiative* shoots up for all rater groups. The percentage of raters focusing on this trait is now above 50 % for all groups. Yet, the group that has the lowest number of raters focusing on *initiative* is actually the group of expert SP-raters (50 %) followed by the group of all SP-raters (56 %). In the group of naïve NS, 58 % of the raters focus on this trait and as many as 78 % of the N2-teachers. When

frequency is considered, the group of the naïve NS tops the list (1.83) followed by the N2-teachers (1.43). The two groups of SP-raters refer less frequently to *initiative* than the other groups when basing their scores on the NORS: all SP-raters refer to it 1.28 times on average, while expert SP-raters only use it 1.17 times. Hence, for the criterion *initiative* H3 is not supported by either scoring method.

When grouping formal traits and communicatively related traits the pattern is as follows. As we can se from Table 15 and Table 16 there are only minor group differences for the formal linguistic traits. The focus on these traits is extensive for all groups. N2-teachers focus less on pronunciation than the other groups, and expert SP-raters use grammar less frequently when scores are based on the NORS as compared to the other groups. Rater training does not seem to have any significant effect on the use of these traits, since the differences between groups are minimal.

When all traits that relate to communicative aspects of speech are grouped together (Table 17 and Table 18) we do however see a positive effect of rater training. In the group of naïve NS only 35 % focus on these traits, compared to 47 % of the N2-teachers, 56 % of all SP-raters and 63 % of the expert SP-raters when scoring impressionistically. The pattern is about the same when frequencies are considered. The group of all SP-raters uses communicative functionality the most (5.19), followed by expert SP-raters (5.00), while N2-teachers (4.31) and naïve NS (4.00) use it somewhat less. When basing their scores on the NORS the pattern is even clearer. Between 66 % and 69 % of raters of all groups focus on these traits, the exception being the group of naïve NS where only 53 % focus on these traits. Expert SP-raters refer to such traits on average 7.50 times, compared to all SP-raters, who use them 7.00 times, N2-teachers 6.78 times and naïve NS only 5.71 times. Hence, while the match between raters' internal criteria and the criteria of the NORS is large for all groups where the formal linguistic traits are concerned, there are indeed quite substantial differences between groups for the trait *communicative ability* when considered separately, as well as for all traits which relate to communicative functionality grouped together. Here, there is a positive correlation between rater training and the match between raters' criteria and the criteria of the NORS as predicted by H3.

So far, we have been looking at the effect of rater training on the criteria raters are supposed to focus on according to the NORS. Some raters do however focus on traits that are not verbalised in the NORS, which would jeopardise validity. For H3 to hold good, raters of Språkprøven should focus less on such traits than the other groups. Two such traits apparent in the data, are

fluency and *content*. When scoring impressionistically, *fluency* is used by a considerable amount of raters of all groups (Table 11): 75 % of the naïve NS refer to it, 61 % of the N2-teachers, and 67 % of all SP-raters and expert SP-raters. When frequency is considered there are only minor differences between the groups, but the group of expert SP-raters does refer a little less often to this trait than the other groups. The group that focuses most on fluency is the one of naïve NS. H3 is therefore sustained for fluency, even though the results are not very convincing.

The pattern is slightly different when the NORS is used as a basis (Table 13 and Table 14). In the group of expert SP-raters there is only a small amount of raters focusing on this trait (33 %). Nevertheless, the group of all SP-raters shows the highest percentage of raters referring to the trait (56 %). So even though the very experienced raters conform to the NORS to a large extent, SP-raters with less experience fail to do so. The group of all SP-raters actually refers to fluency most frequently of all groups (0.90), followed by the group of naïve NS (0.75). Expert SP-raters refer to it 0.67 times on average as compared to 0.52 times by the group of N2-teachers. The differences between the groups are not large, though. H3 is here supported for the expert SP raters, but not for the group of all SP-raters.

Another trait used by many raters despite the fact that it is not one of the criteria of the NORS is *content*. When scoring impressionistically 48 % of the N2-teachers and 28 % of all SP-raters focus on this trait (Table 11 and Table 12). In the group of expert SP-raters, 17 % use it as compared to 8 % of the naïve NS. The N2-teachers also use *content* more frequently than the other groups, referring to it on average 0.83 times. The group of all SP-raters use it 0.49 times but expert SP-raters only 0.17 times.

When scoring impressionistically N2-teachers focus on content to a large extent. When basing their scores on the NORS, they nevertheless manage to conform to the scale to some degree. Only 35 % of the raters of this group refer to content when they use the NORS. This is almost as few as the group of all SP-raters (33 %). N2-teachers do however use the trait somewhat more often than the latter group, 0.57 times as opposed to 0.44 times. In the groups of naïve NS and expert SP-raters, no raters focus on this trait and there is therefore a total correlation between their evaluation and the NORS for this trait. For content, there is some support for H3 as one of the groups of informants without rater training uses it considerably more than the group of expert SP-raters. The fact that SP-raters with less experience also use it, and that naïve NS to a little extent focus on it, does however blur the picture some.

10.3.2 Testing of H4: The effect of rating scale on construct validity.

Hypothesis 4 regards the effect of using a rating scale on construct validity and states that:

- H4: The use of an explicit **rating scale (**NORS) affects **construct validity** (as defined in H3) positively. There is a greater match between the criteria of the NORS and those of the raters when raters base their scores on the NORS than when scoring impressionistically.

H4 is tested against the results presented in Table 21, Table 22, Table 23, Table 24, Table 25, Table 26, Table 27 and Table 28. When investigating the effect of rating scale on scores, group differences are subordinate to differences between scoring methods.

We shall start by looking at the focus on formal traits across scoring methods. As earlier stated, raters of all groups do focus on *grammar*, *vocabulary* and *pronunciation* to a large extent both when scoring impressionistically and when they base their scores on the scale. The group of experienced raters does however show a decline in use of grammar of 17 % from impressionistic to NORS-based scoring (Table 23). The frequency calculation reveals a decline of about 3.00 times in use for all rater groups from impressionistic to NORS-based scorings for this trait. *Pronunciation* and *vocabulary* stay quite constant from impressionistic to NORS-based scorings showing only a moderate decline for all groups with a few exceptions (N2-teachers use of *pronunciation*, naïve NS use of *vocabulary*). In sum, raters focus less on grammar when using the NORS than when scoring impressionistically, while the focus on pronunciation and vocabulary stay relatively constant across scoring methods.

The principal criterion of the NORS; *communicative ability,* is used by the exact same number of raters in the group of expert SP-raters from impressionistic to scale-based scoring. All the other rater groups see an increase in the number of raters focusing on this trait when using the NORS. N2-teachers' use rises the most, by 22 %, followed by all SP-raters (18 %) and naïve NS (16 %). When frequencies are considered, all groups, including the expert SP-raters, use this trait more often in their NORS-based scorings than when scoring impressionistically. Expert SP-raters increase the least (0.33) as opposed to all SP-raters (0.46), N2-teachers (0.48) and naïve NS (0.50). H4 is supported by these data: The NORS emphasises focus on communicative skills, and when raters use the scale, they focus more on this trait than when scoring impressionistically, as predicted by the hypothesis.

Initiative, the other main criterion of the NORS, also sees a considerable increase from impressionistic to NORS-based scoring for all groups. This trait sees an increase in the amount

of raters focusing on it from 35 % (all SP-raters) to 58 % (naïve NS) (Table 23). The frequency table shows similar results (Table 24). All groups use *initiative* more frequently when basing their scores on the NORS. As expected, the group of naïve NS manifests the greatest rise in use (+ 1.83) as compared to the other groups which refer to it little more than one time on average (from 1.00 to 1.17). For initiative, then, H4 is supported.

When all the communicatively related criteria are grouped together there is a considerable increase for all groups, both when percentages of raters (Table 25) and frequency of use (Table 26) are considered. As shown in Table 17 a large number of expert SP-raters referred to communicative functionality when scoring impressionistically (63 %). It is therefore not surprising that this group only sees a minor increase of 4 % from one scoring method to the other, while the group of naïve NS and N2-teachers, who did not focus much on these traits in their impressionistic evaluation increase in their use by 18 % and 22 % respectively. All SP-raters' use rises by only 10 %, however. All groups see an increase in use of between 1.50 and 2.50 which is in support of H4.

Unlike the table of percentages, the results of the frequency analysis show that the group manifesting the greatest increase from impressionistic to NORS-based scoring is actually the group of expert SP-raters. So even though a large amount of raters within this group focus on communicative functionality when scoring impressionistically, they do so to a much larger extent when basing their scores on the NORS. Naïve NS showed least increase when it comes to frequency, even though this group showed the greatest increase in terms of percentage. This is also in favour of H4 which predicts an increased focus on the traits of the NORS when using this scale as a basis for scores.

It remains to see whether the use of the NORS influences raters' use of the criteria *fluency* and *content,* which, as mentioned, are used by raters even though they are not included in the rating scale. In accordance with the rating scale, *fluency* drops quite dramatically for all rater groups from impressionistic to NORS-based scoring (Table 23 and Table 24). The group of expert SP-raters demonstrates the greatest decline with 34 %. A similar decline is found in the group of naïve NS, which drops by 33 %. In the group of N2-teachers 33 % fewer raters focus on *fluency* when basing their scores on the NORS, as opposed to the modest decline of 11 % in the group of N2-teachers. When frequency of use is considered, here too *fluency* is less used when raters draw on the NORS as their basis. There are no major differences between the groups, though, which

decline from 0.28 to 0.61, all SP-raters showing the smallest and N2-teachers the greatest differences between scoring methods. Again H4 is sustained.

Content loses ground in most groups when raters base their scores on the NORS. It is utilised by 17 % fewer raters in the expert SP-rater group, 13 % fewer in the L2-teacher group and 8 % fewer in the naïve NS group. The group of N2-teachers, on the contrary, sees a minor increase of 5 % from impressionistic to scale based evaluation. The variation in frequency from one scoring method to the other is, however, very limited (from 0.00 to - 0.26). The results for most groups are in support of H4.

Summing up the results of the qualitative analysis in relation to H3 and H4, these were the main findings. Both hypotheses are to a large extent supported by the qualitative data. The results show that both rater training and the use of the rating scale have a positive effect on construct validity, defined as the match between the criteria raters use and the criteria of the scale. Trained raters evaluate more in accordance with the scale when scoring impressionistically than do informants without rater training and experience. This is particularly evident in raters' focus on communicative functionality. Focus on *communicative aspects of speech* is primordial in the NORS and is only used to a limited degree by naïve NS and N2-teachers when scoring impressionistically. All SP-raters, and expert SP-raters in particular, focus on the communicative aspects of speech also when scoring on impressionistic grounds. The traits *initiative* and *content,* which are not mentioned in the NORS, but which are still used by some raters, are used to a somewhat larger extent by naïve NS and N2-teachers than by the raters of Språkprøven, but there are only minor differences here. H3 is therefore largely sustained by the data: the more experienced the raters, the more they manage to conform to the NORS and focus not only on formal linguistic traits but on the communicative aspects of speech as well. They also manage to exclude aspects of speech which are not part of the NORS, such as content and fluency.

For H4, which postulates a positive effect of using a rating on construct validity, a similar conclusion may be drawn: i.e. the NORS focuses on communicative functionality, and even in the level descriptors of formal traits, there is an explicit linking to whether or not errors hinder communication. When the informants base their scores on the NORS, there is a rather dramatic rise in the use of the communicatively related criteria for all groups. Both *communicative ability* and *initiative* also see a considerable increase in use. The traits that are not mentioned in the scale, i.e. *fluency* and *content*, decrease in use from impressionistic to NORS-based scorings as predicted by H4.

CHAPTER 11: DISCUSSION

In this chapter I will discuss the results of the study presented in Chapters 9 and 10. Some of the results were as predicted by the hypotheses. Other findings were however surprising and need an explanation. The focus will primarily be on the unexpected results.

Hypotheses H1 and H2 postulate a positive effect of rater training and rating scales on IRR. H1 (the effect of rater training) was supported by the data for the impressionistic scoring but not when raters used the NORS: when scoring impressionistically, trained raters showed a greater degree of internal agreement about the scores than informants without rater training, and there were quite considerable differences between the groups (α= .12 for the N2-teachers and α= .69 for the group of expert SP-raters). When ratings were based on the NORS, however, the effect of rater training seemed to be almost eliminated: there were only minor differences between the groups, and the group that showed the highest estimates of reliability was actually the group of naïve NS. For this scoring method H1 was not supported.

H2 (the effect of rating scale) was supported by the data for the groups of informants without rater training (naïve NS and N2-teachers) but not for the two groups of SP-raters which both scored more reliably when scoring impressionistically than when using the NORS. As mentioned, the group of all SP-raters sees only a very small decline from impressionistic to scale-based scoring, and I would claim their coefficients stay close to constant (from .38 to .36). The group of expert SP-raters however sees a considerable decline (from .69 to.45). Some results were as expected, and some were more surprising and need an explanation. How can we explain that:

- rater training does not have an effect when raters use the NORS?
- expert SP-raters are less agreed when using the NORS than when scoring impressionistically?

Maybe the answers to these questions are to be found in the criteria raters use? In Chapter 10, Section 10.1, differences in raters' use of criteria were grouped in three categories.

- the number of criteria used
- the internal agreement about the criteria used
- the focus on some criteria over other

These same categories will here be used in an attempt to explain inter-rater reliability or the lack of such. I assume that the use of *a limited number of criteria* promotes rater agreement while the use of a large set of criteria hinders this. Moreover, I assume that *internal agreement* about criteria affects rater reliability positively. And finally, I assert that it is easier to yield reliable scores if

raters focus on aspects of speech that relate to a norm of correctness as opposed to traits that cannot be related to such a norm. In Section 11.1 the assumptions are used in a tentative explanation for why some rater groups obtain higher reliability estimates than other groups. In Section 11.2 the assumptions are used to shed light on the question why raters without rater training (naïve NS and N2-teachers) obtain higher estimates of reliability when using the NORS, while this is not the case for the two groups of raters of Språkprøven which both obtain higher estimates of reliability when scoring impressionistically than when using the NORS.[19]

11.1 A tentative explanation of the results of H1.

The comparison between four different groups of informants in Chapter 9 revealed that there are indeed differences between groups as to the internal agreement or inter-rater reliability, and that rater training does have a positive effect on rater-reliability. The results were positive for the impressionistic scoring, but when the informants used the NORS as a basis for their rating, group differences were almost eliminated. The purpose of this section is to search for explanations to the question concerning why trained raters are more in agreement about the scores than the other groups when rating impressionistically, as well as why the effect of rater training seems to be eliminated when raters use the NORS.

11.1.1 H1 and the number of criteria used

It seems intuitively plausible that the more criteria raters use in their evaluation, the harder it is for them to reach an agreement. If raters were to focus on one trait only, the reliability of scores would probably be quite high. When a range of many different criteria are put to use, on the other hand, several sources of disagreement are introduced: not only are raters to agree on a definition of a set of different criteria, but they also have to agree about the weighting of one criterion versus another. If this assumption is correct, rater groups which use few criteria in their evaluation should reach the highest estimates of inter-rater reliability, while the use of many criteria could be a possible explanation for low reliability coefficients.

The assumption is checked against the percentage calculation of raters focusing on different traits for both scoring methods. The issue here is the number of criteria used by raters of each group. Percentages of raters seems more adequate for testing this assumption than does the calculation of how often each trait is used by raters of different groups.

[19] Again, we should keep in mind that there were only minor differences between scoring methods for the group of all SP-raters. This group will therefore not be in focus here.

When scoring impressionistically there is a positive correlation between raters' training and inter-rater reliability as discussed in Chapter 9. The group of expert SP-raters obtains reliability coefficients of α= .69 as compared to the group of raters without rater experience that obtains α= .23 (naïve NS) and α= .12 (N2-teachers).

In the following, I compare only the two extreme groups, that is the groups that obtain the top and bottom scores from the reliability analysis, expert SP-raters and N2-teachers.

Table 29 Ten criteria, extreme groups, impressionistic scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
N2-teachers (α= .12)	96 %	74 %	96 %	61 %	74 %	74 %	52 %	**22 %**	13 %	**48 %**
Expert SP-raters (α= .69)	**100 %**	**100 %**	**100 %**	**67 %**	**83 %**	**83 %**	**83 %**	0 %	**67 %**	17 %

The group of N2-teachers uses all of the ten criteria in their evaluation while the group of expert SP-raters does not focus on *initiative* and consequently uses only nine of the ten criteria. Even though the group of N2-teachers applies more criteria in their evaluation than does the group of expert SP-raters this is not the whole picture. Just as relevant is the question of the number of raters focusing on the different aspects. When these facts are taken into account a different pattern is visible. In the group of expert SP-raters many criteria are used by a great number of raters. Three criteria are used by all raters of this group and three more are used by 83 % of the raters. In other words, six of the ten criteria are used by more than 80 % of the raters of this group. When the two groups are compared for the number of raters focusing on different criteria the group of expert SP-raters outdoes the N2-teachers for eight of the ten traits. There therefore seems to be a greater spread across the range of criteria in the group of expert SP-raters than in the group of N2-teachers. And yet, the expert SP-raters obtain higher reliability estimates than the group of N2-teachers. The conclusion based on the impressionistic scoring then, must be that the use of a large number of criteria is not a plausible explanation for low inter-rater reliability estimates. There is no negative correlation to be found between the number of criteria used and low reliability estimates in these data.

Let us now turn to the NORS-based ratings to see whether the results are the same. The results of the reliability study for this scoring method were as follows: the group of naïve NS obtained the highest estimates of reliability with α= .47 and the group of all SP-raters obtained the lowest

with α= .36. The differences between the groups are however smaller than the differences when raters scored impressionistically. Again the two extreme groups will be compared.

Table 30 Ten criteria, extreme groups, NORS-based scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Naïve NS (α= .47)	**100 %**	92 %	**100 %**	42 %	58 %	75 %	67 %	**58 %**	8 %	0 %
All SP-raters (α= .36)	95 %	92 %	95 %	**56 %**	**95 %**	**80 %**	**69 %**	56 %	**28 %**	**33 %**

Naïve NS use nine of the ten criteria, while all SP-raters use all of the ten criteria in their evaluation. In the group of naïve NS three traits are used by more than 80 % of the raters as compared to five criteria in the group of all SP-raters. Six of the ten criteria are used by a larger number of all SP-raters than naïve NS, one trait is used by the same number of raters in the two groups and three criteria are used by a larger number of naïve NS than all SP-raters. Hence, there is a greater spread across criteria for the group of all SP-raters than for the group of naïve NS. When scores are based on the NORS, then, the assumption that a great spread across criteria may have a negative effect on inter-rater reliability is sustained.

11.1.2. H1 and the internal agreement about criteria

A second possible explanation for rater reliability is the internal agreement between raters about the criteria upon which they base their scores. If raters of a group were totally agreed with each other when it comes to which traits to emphasise, this should affect inter-rater reliability positively, and if raters focus on different aspects of performance in their ratings, this, should lead to a greater disagreement between raters about the scores they set. Again, the interesting groups for comparison are those which obtain the highest versus the lowest estimates of inter-rater reliability. High internal agreement between raters of one group is manifested if 100 % or 0 % of the raters focus on a certain trait.

Table 31 Ten criteria, extreme groups, impressionistic scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Comu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
N2-teachers (α= .12)	**96 %**	74 %	**96 %**	61 %	74 %	74 %	52 %	22 %	**13 %**	48 %
Expert SP-raters (α= .69)	**100 %**	**100 %**	**100 %**	67 %	**83 %**	**83 %**	**83 %**	**0 %**	67 %	**17 %**

This is a copy of Table 29 repeated here for the sake of convenience

As evident in Table 31 there is a much greater degree of internal agreement about the criteria amongst the expert SP-raters than amongst the N2-teachers. For as many as four of the traits there is an absolute agreement between the expert SP-raters, that is, either all or no raters focus on these traits. Three more traits are used by more than 80 % of the raters, which means that most of the raters of this group are agreed. One trait is only used by 17 %, so there is strong agreement for this trait as well. There is considerable disagreement (67 %) for only two of the ten criteria for the group of expert SP-raters when scoring impressionistically.

N2-teachers, on the other hand, do not concur much about the criteria of speech. In fact, there is not full agreement for any of the ten traits. Only two traits are used by more than 80 % and one trait is applied by 13 % of the raters. This is the closest this group comes to an agreement. For five traits there is extensive disagreement between raters, that is, about half of the raters focus on the trait while the other half fail to do so (48 % to 74 %).

The assumption about a positive correlation between internal agreement about the criteria and high IRR estimates is sustained by the data for the impressionistic scoring method. The expert SP-raters, who obtain the highest estimates of reliability, are to a very great extent agreed about the criteria of speech.

We shall now turn to a comparison of extreme groups for the NORS-based scoring method.

Table 32 Ten criteria, extreme groups, NORS-based scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Naïve NS (α=.47)	**100 %**	**92 %**	**100 %**	42 %	58 %	75 %	67 %	58 %	8 %	**0 %**
All SP-raters (α= .38)	**95 %**	**92 %**	**95 %**	56 %	**95 %**	**80 %**	69 %	56 %	28 %	33 %

This is a copy of Table 30, repeated here for the sake of convenience.

The group of naïve NS shows total agreement for three of the ten traits. Two traits are used by all raters of this group, while one trait is not utilised by any raters. In addition one trait is used by as many as 92 % and another by as few as 8 %. For five traits there is disagreement between raters (from 42 % to 75 %).

For the whole group of SP-raters, on the other hand, there is not full agreement for any traits. Three traits are put to use by 95 % of the raters, though, and two more by more than 80 %, meaning that the majority of the raters are agreed about these traits. Five traits are used to a varying extent by raters of this group (from 28 % to 69 %).

The conclusion, then, is the same as the one drawn for the impressionistic data. There does indeed seem to be a positive correlation between the internal agreement about the criteria of speech and high inter-rater reliability estimates. Hence, internal agreement about the definition of the test construct seems to be a prerequisite for high inter-rater reliability.

11.1.3. H1 and the focus on formal linguistic traits over communicative functionality

The final assumption about the relation between the use of criteria and reliability made in Section 11.1 regards the focus on some criteria over others. The assumption made is that it is easier to agree about the score when based on traits that refer to a norm of correctness (formal linguistic traits: *grammar, pronunciation*, and *vocabulary*) than on traits that lack reference to such a norm (communicative functionality: *communicative ability, comprehension, intelligibility, initiative,* and *strategies*). *Content* is here classified together with communicative functionality, not because it is theoretically convincing to do so, but rather because this trait shares an important characteristic with the traits classified under communicative functionality: Like these traits, it cannot be related to a norm of correctness, it has to be scored based on raters' subjective evaluation of quality.

In the investigation of this assumption both percentages of raters and frequency of use are relevant. Tables displaying the use of all criteria separately as well as those grouping formal linguistic traits and communicative functionality are used as a basis for this investigation.

Table 33 Ten criteria, extreme groups, impressionistic scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
N2-teachers (α= .12)	96 %	74 %	96 %	74 %	74 %	52 %	**22 %**	13 %	**48 %**
Expert SP-raters (α= .69)	**100 %**	**100 %**	**100 %**	**83 %**	**83 %**	**83 %**	0 %	**67 %**	17 %

This is a copy of Table 29 repeated here for the sake of convenience

For the impressionistic scoring, expert SP-raters obtain the highest IRR estimates of the four rater groups. Do they focus more on formal linguistic traits and less on communicative functionality as assumed? As evident in Table 33 above, the largest difference between the two extreme groups when it comes to formal linguistic traits, is the lack of focus on *pronunciation* by the group of N2-teachers. In this group somewhat fewer raters focus on *grammar* and *vocabulary* than in the group of expert SP-raters. Hence, these traits, which are assumed to correlate positively with rater reliability, are put to use by a larger number of expert SP-raters than by N2-teachers. This supports the assumption.

However, communicative functionality is also used by more expert SP-raters than N2-teachers. Four of the five communicatively related traits are used by a larger number of expert SP-raters, the exception being *initiative* which is used by 22 % of the N2-teachers as opposed to 0 % of the expert SP-raters. *Content* is used by quite a large number of N2-teachers (48 %) while only a limited number of expert SP-raters focus on this trait (17 %).

Table 34 Ten criteria, extreme groups, impressionistic scoring, frequencies.

Rater groups	grammar	pronun-ciation	vocabu-lary	commu-nicative ability	intelligibi-lity	compre-hension	initiative	Strategies	**content**
N2-teachers (α= .12)	3.96	1.61	3.17	**2.09**	1.35	0.87	**0.30**	0.22	**0.83**
expert SP-raters (α= .69)	**4.17**	**3.67**	**3.50**	1.50	**1.83**	**1.67**	0.00	**0.67**	0.17

The results of the frequency calculation match those of the percentage calculations to a large degree. The only difference is that while the number of raters focusing on *communicative ability* was higher in the group of expert SP-raters, the group of N2-teachers actually refers more often to this trait than do the experts. This may be due to the fact that expert SP-raters specify communicative ability to a greater extent than N2-teachers, as evident from the fact that they use

three of the four other communicatively related traits to a larger extent than do the N2-teachers. N2-teachers however refer more frequently to *content* than the expert SP-raters.

The table grouping formal linguistic traits and communicative functionality together should not imply any great surprises:

Table 35 Formal traits versus communicative functionality, extreme groups, impressionistic scoring, percentages and frequencies.

	formal linguistic traits		**communicative functionality and content**	
	percentages	frequencies	percentages	frequencies
N2-teachers (α=.12)	89 %	2.91	47 %	4.31
expert SP-raters (α= .69)	**100 %**	**3.38**	**63 %**	**5.00**

For the formal traits, the group of expert SP-raters outperforms the group of N2-teachers both when percentages of raters and frequency of use are considered. The interesting point is that this goes for communicative functionality as well. And yet, they agree more about the scores. For the impressionistic scoring, then, there is no support for the assumption that focus on formal correctness has a positive effect while focus on communicative functionality has a negative effect on inter-rater reliability. Hence, the assumption is not sustained.

Table 36 Ten criteria, extreme groups, NORS-based scoring, percentages.

Rater groups	Grammar	Pronun-ciation	Vocabu-lary	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Naïve NS (α= .47)	**100 %**	**92 %**	**100 %**	58 %	75 %	67 %	**58 %**	8 %	0 %
All SP-raters (α=. 38)	95 %	**92 %**	95 %	**95 %**	**80 %**	**69 %**	56 %	**28 %**	**33 %**

This is a copy of Table 30, repeated here for the sake of concvenience.

When basing their scores on the NORS, there are indeed differences between the extreme groups as to the criteria they use. A larger number of naïve NS focus on the formal traits *grammar* and *vocabulary*. For *pronunciation* the percentages are the same for the two groups. In the naïve NS group fewer raters focus on communicative functionality, the only exception being *initiative* which is used by 2 % more raters of this group than by the group of all SP-raters. For the NORS-based scoring, then, the assumption is supported by the data.

The pattern is however less clear when frequencies are considered.

Table 37 Ten criteria, extreme groups,, NORS-based scoring, frequencies

Rater groups	grammar	pronun-ciation	vocabu-lary	commu-nicative ability	intelligibi-lity	compre-hension	initiative	Strategies	content
Naïve NS (α= .47)	1.00	**2.92**	**4.00**	1.50	**2.42**	**1.42**	**1.83**	0.8	0.08
all SP-raters (α=. 36)	**1.05**	2.31	3.10	**2.72**	1.46	1.10	1.28	**0.44**	**0.44**

The group of naïve NS refers more frequently to *pronunciation* and *vocabulary* than does the group of all SP-raters. For *grammar* there are only minor differences between the groups, yet the SP-raters use it the most. For communicative functionality, the pattern is blurry in that all SP-raters refer most often to the traits *communicative ability* and *strategies,* but the group of naïve NS uses *intelligibility, comprehension* and *initiative* more frequently than all SP-raters. The SP-raters group also refers more frequently to *content* than does the group of naïve NS.

Table 38 Formal traits versus communicative functionality, extreme groups, NORS-based scoring, percentages and frequencies.

	formal linguistic traits		**communicative functionality**	
	percentages	frequencies	percentages	Frequencies
naïve NS (α= .47)	**97 %**	**2.64**	53 %	5.71
all SP-raters (α=. 30)	94 %	2.15	**66 %**	**7.00**

When grouping the traits into two main categories, a clearer pattern is visible: the naïve NS refer more often to formal linguistic traits and less to communicative functionality than does the group of all SP-raters. The assumption is indeed supported by the data for the NORS-based scoring. That naïve NS' focus on formal traits, and the SP-raters' focus on communicative functionality may therefore be one possible explanation for the surprising fact that the group of naïve NS obtain higher estimates of reliability than the group of all SP-raters when scores are based on the NORS.

Summing up the results presented in Section 11.1.3, we see that the assumption is sustained for the NORS-based but not for the impressionistic scoring method. When raters base their scores on the rating scale of Språkprøven there is a positive correlation between the focus on formal traits and inter-rater reliability: i.e. raters' focusing on norm-related formal linguistic traits affects

rater-reliability positively, while the opposite seems to be the case for communicative functionality, as well as for *content:* the group that focuses the most on these traits obtains lower estimates of inter-rater reliability.

For the impressionistic scoring, however, a different pattern comes into view. The group of expert SP-raters focuses more extensively on communicative functionality than the group of N2-teachers with which it is compared. Yet, expert SP-raters obtain considerably higher inter-rater coefficients. For this scoring method, the assumption is not sustained.

These results may be interpreted in the following way: the rating of communicative functionality is difficult because it does not relate to a norm of correctness. The group of naïve NS focuses almost exclusively on the formal linguistic traits when scoring impressionistically and therefore manages to agree about which scores to assign. SP-raters with mixed rater-experience focus more on the communicative aspects of speech than do the groups of naïve NS and N2-teachers without rater training, but by doing this they jeopardise inter-rater reliability. The only group that manages to focus on the communicatively functionality and still obtain high inter-rater reliability estimates is the group of expert SP-raters. In order to be able to focus on these less norm-referenced aspects of speech and still be in agreement with the other raters, it seems that extensive rater training and experience is necessary.

Summing up the discussion of Section 11.1.1 in relation to the effect of rater training on inter-rater reliability as formulated in H1, these are the main points. Three assumptions were postulated about possible differences between rater groups affecting inter-rater reliability. The first assumption was that rater groups obtaining the highest estimates of reliability focus on a more limited number of criteria than do rater groups with lower reliability estimates. This assumption was not sustained by the data for the impressionistic scoring method, but sustained when the NORS was used as a basis.

The second assumption was that internal agreement between raters about the criteria of speech would have a positive effect on inter-rater reliability. This assumption was sustained by the data for both impressionistic and the NORS-based scoring methods: in both cases, rater groups that are most internally agreed about the criteria obtain the highest estimates of reliability.

A final assumption was made regarding the focus on certain criteria over others. A claim was made that focus on formal linguistic traits which to a large extent refer to a norm of correctness, would promote reliability, while the focus on communicatively functionality would affect reliability negatively. This assumption was supported by the data when raters based their scores on the NORS, but not when they scored on impressionistic grounds.

Consequently, for the NORS-based scoring method all three assumptions were sustained by the data. When scores were given on impressionistic grounds, on the other hand, the assumption that agreement about the criteria promotes inter-rater reliability was the only one to find support in the data. There are no obvious explanations for these differences, however, one possible reason relates to the groups that obtain the highest estimates of reliability for the two scoring methods; the expert SP-raters for the impressionistic and the naïve NS for the NORS-based scoring. The results of the three assumptions presented in this section, suggest that there are different explanations for rater agreement about the scores for the two groups of informants. For naïve NS to score reliably it seems to be a condition that they focus on a limited set of criteria, that they use the same criteria, and finally that these criteria relate to a norm of correctness. The group of expert SP-raters with extensive training and experience with the NORS, however, seems to be less restricted by such conditions. They manage to obtain the highest level of inter-rater reliability, even though they focus on a wide range of criteria in their ratings, and despite the fact that these criteria are not related to a norm of correctness. The expert SP-raters are not only agreed about the scores they give, but in addition, they agree about the underlying construct of the test. Indeed, they do, to a large extent, focus on the same aspects of performance.

11.2 A tentative explanation of the results of H2: the effect of rating scale on inter-rater reliability

H2 postulates a positive effect of rating scale (NORS) on IRR. H2 was supported by the data for all rater groups except for the group of expert SP-raters, which showed considerably higher IRR estimates when scoring impressionistically than when using the NORS, and the group of all SP-raters, which showed a minor decrease from impressionistic to scale-based scoring.

The purpose of this section is to use the qualitative data in a quest for explanations to these differences in reliability estimates from impressionistic to NORS-based scoring. The three assumptions presented in the introduction of the present chapter and applied in Section 11.1 are once more used as a basis.

11.2.1 H2 and the number of criteria used

The assumption that the use of a large number of criteria affects inter-rater reliability negatively was raised in Section 11.1.1. If this assumption is correct, the groups of naïve NS and N2-teachers should focus on more criteria when scoring impressionistically, while the opposite

should be the case for the two groups of SP-raters. In order to investigate this assumption, we should study each rater group separately. Yet, it may be interesting to have a quick look at the results of all raters grouped together first.

Table 39 Joint table for all informants, differences between scoring methods, percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Imressionis-tically	99 %	88 %	93 %	66 %	70 %	80 %	55 %	18 %	24 %	31 %
With NORS	96 %	88 %	96 %	49 %	89 %	78 %	65 %	64 %	27 %	28 %
Difference	- 3%	-	+ 3 %	- 17 %	+ 19 %	- 2 %	+ 10 %	+ 46 %	+ 3 %	- 3 %

This is a copy of Table 19 presented in Chapter 10 and repeated here for the sake of convenience.

As Table 39 shows there is a greater spread of criteria when raters use the NORS than when scoring impressionistically as evident from the last row displaying the differences between the percentages of raters focusing on distinct traits from one scoring method to the other. Five traits are used by a larger number of raters when scores are based on the NORS, one is used by just as many raters across scoring methods and only four criteria are used by more raters when scoring impressionistically. In addition, the criteria that manifest an increase in use from impressionistic to scale based scoring do so to a considerable extent. The positive numbers of the last row total 81 (%) while the corresponding sum of the negative numbers is only 25 (%). This means that even though the number of criteria used more frequently by one scoring method over the other is not very large, the number of raters focusing on certain criteria is much higher when scores are based on the NORS. It remains to be seen whether there are differences between the groups in the number of criteria used from impressionistic to NORS-based scorings.

Table 40 Difference between scoring methods, ten criteria, all rater groups, percentages, increase and decrease indexes.

Rater groups	Gram-mar	Pronun-ciation	Voca-bulary	Fluency	Commu-nicative ability	Intel-ligibi-lity	Compre-hension	Initia-tive	Strategies	Content	**Increase index +**	**Decrease index -**
Naïve NS	-	-8 %	**+8 %**	- 33 %	**+ 16 %**	-	**+ 9 %**	**+ 58 %**	**+ 8 %**	- 8 %	**99**	**49**
N2-teachers	-	**+4 %**	-	- 22 %	**+ 22 %**	**+2 %**	**+ 5 %**	**+ 56 %**	**+ 22 %**	- 13 %	**111**	**35**
All SP-raters	-5 %	-	**+3 %**	- 11 %	**+ 18 %**	-5 %	**+ 13 %**	**+ 35 %**	- 11 %	**+ 5 %**	**74**	**32**
Expert SP-raters	-17 %	-	-	- 34 %	-	**+17 %**	- 16 %	**+ 50 %**	- 34 %	- 17 %	**67**	**118**
										Total increase and decrease indexes from impressionistic to NORS-based scorings:	**351**	**234**

Increase index refers to the sum of all positive numbers for each rater group, i.e. the sum of all numbers indicating an increase in use from impressionistic to NORS-based scoring. Decrease index is the sum of all negative numbers, indicating that the criteria are used by a larger number of raters when they score impressionistically than when basing their scores on the NORS.

The numbers indicating increase and decrease in the use of criteria from impressionistic to NORS-based methods are summed up for each rater group and presented in the two columns to the right of the table under the labels "increase index" and "decrease index". Strikingly enough, the total increase index is considerably larger than the total decrease index. In other words, a larger percent of raters use the traits when scoring with the NORS than without it.

If we look at each group a part, we see that in the group of naïve NS, a larger number of raters focus on the different traits when scorings are based on the NORS. The same is the case for the group of N2-teachers. These groups show higher degrees of IRR when using the NORS, hence the assumption that a limited number of criteria promotes IRR is not sustained for these groups.

The group of all SP also shows an increase in raters using the distinct traits when scorings are based on the NORS. But as distinct from the two groups of non-raters, this group shows higher estimates of reliability when scoring impressionistically than NORS-based. Hence, for this group the assumption is sustained.

The expert group is a case apart. Not only does this group stand out from the other groups in that it obtains considerably higher estimates of inter-rater reliability when scoring impressionistically than when using the NORS. In addition this group differs from the others in that the decrease index is larger than the increase index. This means that a larger number of raters use the distinct criteria when scoring impressionistically than when they use the NORS. The assumption is therefor falsified for all groups, except for the group of all SP-raters for which it is supported.

11.2.2. H2 and internal agreement about the criteria

We shall now shift the focus from the number of criteria used to the internal agreement between raters about the criteria. The second assumption presented in Section 11.1.2 was that agreement between raters of one group as to which criteria to use would affect reliability positively. The assumption was sustained when the focus was on differences between rater groups. Indeed, the groups that showed the highest estimates of reliability were more in agreement about the criteria than the other groups. We shall now see whether this assumption may explain differences in reliability estimates from one scoring method to another as well. Again, the assumption is tested against the results of separate groups, and the tables presenting all raters together are used mostly as background information.

Table 41 Ten criteria, all raters (n= 74), percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Fluency	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Imressionis-tically	**99 %**	**88 %**	**93 %**	66 %	70 %	**80 %**	55 %	18 %	24 %	31 %
With NORS	**96 %**	**88 %**	**96 %**	49 %	**89 %**	78 %	65 %	64 %	27 %	28 %

When all raters are grouped together no trait obtains full agreement (0 % or 100 %) in either scoring method. There are no major differences to be found from impressionistic to NORS-based scoring where agreement about criteria is concerned. This may perhaps be explained by the heterogeneity of the group when all raters are lumped together. The results for each group separately is presented below.

Table 42 Ten criteria, all rater groups, both scoring methods, percentages, totals of agreement.

Rater groups	Scoring method/ IRR	Gram-mar	Pronun-ciation	Voca-bulary	Fluency	Com. ability	Intelli-gibility	Com-prehensi on	Initi-ative	Strategie s	Content	Full agree-ment	High agree-ment	TOTAL of full and high agreem.
Naïve NS	IMPR. (.24)	**100 %**	**100 %**	**92 %**	75 %	42 %	75 %	58 %	**0 %**	**0 %**	**8 %**	4	2	6
	NORS (.47)	**100 %**	**92 %**	**100 %**	42 %	58 %	75 %	67 %	58 %	**8 %**	**0 %**	3	2	5
N2-teachers	IMPR. (.12)	**96 %**	74 %	**96 %**	61 %	74 %	74 %	52 %	22 %	**13 %**	48 %	0	3	3
	NORS (.40)	**96 %**	78 %	**96 %**	39 %	**96 %**	78 %	57 %	78 %	35 %	35 %	0	3	3
All SP-raters	IMPR. (.38)	**100 %**	**92 %**	**92 %**	67 %	77 %	**85 %**	56 %	21 %	39 %	28 %	1	3	4
	NORS. (.36)	**95 %**	**92 %**	**95 %**	56 %	**95 %**	**80 %**	69 %	56 %	28 %	33 %	0	5	5
Expert SP-raters	IMPR. (.69)	**100 %**	**100 %**	**100 %**	67 %	**83 %**	**83 %**	**83 %**	**0 %**	67 %	**17 %**	4	4	8
	NORS. (.45)	**83 %**	**100 %**	**100 %**	33 %	**83 %**	**100 %**	67 %	50 %	33 %	**0 %**	4	2	6
Full agree-ment		4	**3**	**3**	0	0	**1**	0	2	1	**2**	16		
High agree-ment		4	**3**	**5**	0	4	**3**	1	0	2	**2**		24	
TOTAL of full and high agrem.		8	**6**	**8**	0	4	**4**	1	2	3	**4**			40

The table presents differences between the two scoring methods for all rater groups. The reliability estimates of each group across scoring methods are presented in the second column. The third to last column display the number of criteria that achieve full agreement by one rater group. The second to last column show the number of criteria that obtain high, yet not full, agreement. High agreement is taken to mean that more than 80 % or fewer than 20 % of the raters focus on a given trait. The last column sums up the traits that obtain full or high agreement for each rater group. The third to last row present the number of raters who achieve full agreement about the criterion in question for both scoring methods. The second to last row show the number of raters obtaining high but not full agreement, and in the last row full and high agreement about the criteria are summed up. The traits that achieve full or high agreement are written in bold letters.

As evident in Table 42 naïve NS show full agreement for four out of ten criteria, and high agreement for two criteria when scoring impressionistically. When using the NORS there is however a small decline in agreement. Now, there is full agreement about three out of ten criteria and high agreement about two criteria. The differences between the scoring methods are not large, though. For this group the assumption about a positive correlation between agreement of criteria and inter-rater reliability is not sustained.

The group of N2-teachers also shows higher reliability estimates when using the NORS, and for the assumption to be sustained, N2-teachers should be more agreed when using the NORS than when scoring impressionistically. However, there is not full agreement for any trait in either scoring method and only three traits obtain degrees of high agreement for both methods in this group. The degrees of agreement are identical across scoring methods, and yet they obtain higher reliability estimates when using the NORS. Consequently, the assumption is once more refuted by the data.

All SP-raters score somewhat less reliably when using the NORS than when scoring impressionistically, and consequently, according to assumption two there should be a larger degree of internal agreement about criteria for the impressionistic scoring for this group. Nevertheless, Table 42 shows a lack of total agreement about any traits for this group when scoring impressionistically, but there is high agreement about four traits. When using the NORS, though, there is a small increase in the internal agreement, indeed there is full agreement about one and high agreement about four other traits. Hence, the assumption is not sustained by the data for this group.

Finally, the group of expert SP-raters shows the highest degree of internal agreement about criteria. This group obtains considerably higher IRR estimates when scoring impressionistically than when using the NORS. Hence, for the assumption to hold water for this group, there should be a higher degree of internal agreement of criteria when scoring impressionistically than when scores are based on the NORS. As displayed in the table, this is indeed the case. When scoring impressionistically, there is full agreement about four criteria and high agreement about another four. When the expert SP-raters base their scores on the NORS on the other hand, the number of criteria that obtains total agreement stays invariable, yet there is a decline of the traits that obtain high agreement from four to two traits. For this group, then, the assumption that agreement about criteria promotes inter-rater reliability is sustained.

Summing up, then, the assumption regarding a positive relation between agreement about criteria and agreement about scores is only sustained for one out of four groups. While this

assumption did indeed seem to have explanatory power when related to the question why some rater groups were more agreed about the scores than others (Section 11.1.2.), it is less powerful as an explanation of differences in reliability estimates by the same rater groups across scoring methods.

11.2.3 H2 and the focus on formal linguistic traits over communicatively related traits.

The third condition assumed to affect inter-rater reliability positively is the focus on formal traits over communicatively related traits. In this section this assumption is applied in an attempt to explain differences in reliability estimates for each group across scoring methods. The purpose is to investigate whether focus on formal versus communicatively related traits may explain why the groups of informants without rater training obtain higher estimates of IRR when basing the scores on the NORS, and, similarly, why both groups of SP-raters obtain lower estimates of reliability when using the NORS. Both percentages and frequency are used as a basis for this investigation. The traits are first studied separately and thereafter grouped in two main categories as in Section 11.1.3 above.

The first two tables show the results for all raters grouped together. These tables were introduced in Chapter 10 and repeated here for the sake of convenience.

Table 43 Difference between scoring methods, ten criteria, all raters (n=74), percentages.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impres-sionistic	99 %	88 %	93 %	70 %	80 %	55 %	18 %	24 %	31 %
NORS-baseds	96 %	88 %	96 %	89 %	78 %	65 %	64 %	27 %	28 %
Diffe-rence	- 3%	-	+ 3 %	+ 19 %	- 2 %	+ 10 %	+ 46 %	+ 3 %	- 3 %

This is a copy of Table 19 presented in Chapter 10 and repeated here for the sake of convenience.

The focus on *formal linguistic traits* stays relatively constant from one scoring method to the other when all raters are grouped together. Raters do not seem to use the formal traits to a larger extent when using the NORS. *Pronunciation* stays constant, *grammar* is used a little less and *vocabulary* sees a small increase from impressionistic to NORS-based scoring.

Communicatively related traits on the other hand, see a large increase for three traits and a moderate increase for one trait, while two traits drop a little in use from impressionistic to

NORS-based scoring. The increase and decrease of formal and communicatively related traits are more apparent in the table below where the traits are grouped together in two main categories:

Table 44 Difference between scoring methods, formal traits and communicative functionality, all raters (n=74), percentages.

Scoring methods	**Formal linguistic traits**	**Communicative functionality**
Impressionistic	93 %	46 %
NORS-based	93 %	59 %
Difference	-	+ 13 %

Based on Table 19.

As we see, *formal linguistic traits* stay constant from impressionistic to NORS-based scoring. The *communicatively related traits* as a group, though, are used by 13 % more raters when basing their scores on the NORS. The various groups may however vary in their focus on certain traits across scoring methods. I will return to this after having investigated the frequency of use of criteria between scoring methods. Again all raters are grouped together.

Table 45 Difference between scoring methods, ten criteria, all raters (n=74), frequencies.

Scoring methods	Grammar	Pronun-ciation	Vocabu-lary	Commu-nicative ability	Intelligibi-lity	Compre-hension	Initiative	Strategies	Content
Impres-sionistic	4.01	2.45	3.31	2.00	1.68	1.07	0.24	0.34	0.53
NORS-based	1.04	2.36	3.14	2.47	1.59	1.09	1.42	0.39	0.41
Diffe-rence	- 2.97	- 0.09	- 0.17	+ 0.47	- 0.09	+ 0.02	+ 1.18	+ 0.05	- 0.12

This is a copy of Table 20 presented in Chapter 10 and repeated here for the sake of convenience.

The results of the frequency investigation match the results shown in Table 43 to a large extent. Not only do the formal linguistic traits fail to increase. They even see quite a considerable decrease in use. This is particularly so for *grammar*, while *pronunciation* and *vocabulary* stay more or less invariable. Four of the six communicatively related traits show an increase in use when scores are based on the NORS. The results are summed up and presented below:

Table 46 Difference between scoring methods, ten criteria, all raters (n=74), frequencies.

Scoring methods	**Formal linguistic traits**	**Communicative functionality**
Impressionistic	3.26	0.98
NORS-based	2.18	1.23
Difference	- 1.08	+ 0.25

Formal linguistic traits decrease in use while the opposite is the case for the communicatively related traits.

Let us now see whether the results are the same when each rater group is studied separately.

Table 47 Difference between scoring methods, all rater groups, percentages, increase and decrease indexes.

	Formal linguistic traits					**Communicative functionality**					**Content**		
Rater groups	grammar	pronuncia-tion	vocabu-lary	**increase index**	**de crease index**	commu-nicative ability	intelligibi-lity	compre-hension	initiative	strategies	content	**increase index**	**decrease index**
Naïve NS	-	-8 %	+8 %	-	-	+ 16 %	-	+ 9 %	+ 58 %	+ 8 %	- 8 %	**+ 83**	-
N2-teachers	-	+4 %	-	**+ 4**	**-**	+ 22 %	+2 %	+ 5 %	+ 56 %	+ 22 %	- 13 %	**+ 94**	-
All SP-raters	-5 %	-	+3 %	**-**	**- 2**	+ 18 %	-5 %	+ 13 %	+ 35 %	- 11 %	+ 5 %	**+ 55**	-
Expert SP-raters	-17 %	-	-	**-**	**- 17**	-	+17 %	- 16 %	+ 50 %	- 34 %	- 17 %	-	-

Table 47 *presents the differences between the two scoring methods across rater groups. The dark grey columns present the total indexes of rise or fall for each group for formal linguistic and communicatively related traits separately.*

For the naïve NS, the formal traits stay stable from one scoring method to the other, while there is a relatively large increase index for communicative functionality. Assumption three cannot explain why raters of this group obtain higher estimates of IRR when using the NORS: Raters focus more on communicative functionality when using the scale, and still, as we see, obtain higher estimates of reliability. N2-teachers show a very modest increase in the formal traits and an even larger increase in the communicatively related traits compared to the naïve NS group. Again, then, the assumption lacks support.

The case is similar for the group of all SP-raters: the formal traits stay close to invariable from impressionistic to NORS-based scoring, while the communicatively related traits see a rather considerable increase in use. As this group shows higher estimates of reliability when scoring impressionistically, the assumption is however supported. We should nevertheless keep in mind the relatively small difference between reliability coefficients from across scoring methods for this group, and interpret the results with care.

The group of expert SP-raters, on the other hand, shows considerably higher reliability estimates when scoring impressionistically than when using the NORS. If assumption 3 holds good, raters of this group should therefore manifest a decrease in the use of formal traits and an increase in focus on communicative functionality from impressionistic to scale based scoring. As apparent in Table 47 this is only partly true. Indeed, the expert SP-raters do focus more on formal linguistic traits when scoring impressionistically than when using the NORS. But the communicatively related traits stay invariable across scoring methods: some traits rise and others fall but summed up the total indexes are zero.

Table 48 below shows the differences in frequency of use from impressionistic to NORS-based scorings across rater groups.

Table 48 Difference between scoring methods, all rater groups, frequencies, increase and decrease indexes.

	Formal linguistic traits					**Communicatively related traits**					**Content**		
Rater groups	grammar	pronuncia-tion	vocabu-lary	**increase index**	**decrease index**	commu-nicative. ability	intelligi-bility	compre-hension	initiative	strategies	content	**increase index**	**decrease index**
Naïve NS	-3.08	-0.16	+0.17	-	- 3.07	+0.50	+0.67	-	+ 1.83	-	-	+ 3.00	-
N2-teachers	-2.92	+0.56	-0.43	-	- 2.79	+0.48	+0.04	+0.04	+1.13	+0.26	-0.26	+ 1.69	-
All SP-raters	-2.98	-0.43	-0.13	-	- 3.54	+0.46	-0.39	+0.02	+1.00	-0.05	-0.05	+ 0.99	-
Expert SP-raters	-3.00	-0.50	-0.33	-	- 3.83	+0.33	+0.84	-0.17	+1.17	-0.34	-0.17	+ 1.66	-

The frequency table shows that all groups focus less on formal linguistic traits and more on communicative functionality when ratings are based on the NORS. This is evident from the columns presenting the increase and decrease indexes in the middle and at the right end of the table.

All groups show a decrease index for the formal traits between - 2.79 and - 3.83. N2-teachers show the smallest degree of decrease, due to the increase in reference to pronunciation from impressionistic to scale based evaluation while the group of expert SP-raters shows the largest decrease index for these traits.

The communicatively related traits increase in use for all groups when using the NORS. The naïve NS rise the most (+ 3.00) and all SP-raters the least (0.99), but on the whole these traits are used more extensively by all raters when basing their scores on the NORS than when scoring impressionistically. It is therefore obvious that the increase in reliability from impressionistic to scale based scoring for the naïve NS and N2-teachers cannot be explained by an increased focus on formal linguistic traits and a similar decrease in focus on communicatively related traits when using the NORS. Assumption three is not sustained by the frequency calculation for these groups.

The two groups of SP-raters show lower estimates of IRR when basing their scores on the NORS than when scoring impressionistically If assumption three were correct, these groups should therefore focus more on formal linguistic traits and less on communicative functionality when scoring impressionistically than when using the NORS. As shown in Table 48 above, this is indeed the case. Like the other groups SP-raters show a considerable decrease in their use of formal traits and an increase in use of communicative functionality when basing their scores on the NORS. For these groups, then, assumption three is sustained by the data when frequency of use is considered.

The explanatory power of the assumptions in explaining differences in IRR across scoring methods, may be summed up as follows: The first assumption (number of criteria), lacked support in the data for all rater group except for the group of all SP-raters. This group sees a minor decrease in reliability from impressionistic to scale-based scoring, and indeed, there is an increase in use of various criteria when scores are based on the NORS. However, these results should be interpreted with care, since the difference of reliability estimates across scoring methods is minimal.

The second assumption (agreement about criteria) was powerful as an explanation of why some rater groups obtained higher estimates of reliability than others, as discussed in Section 11.1.2. It is however less suitable to explain why the same rater groups obtain different IRR estimates across scoring methods. Here it is only sustained for one out of four groups, that is the expert SP-raters.

The results were also inconclusive for the third assumption (focus on formal over communicatively related traits). There was a positive correlation for some groups but not for others. The two groups of informants with no rater training (naïve NS and N2-teachers) showed higher IRR estimates when using the NORS than when scoring impressionistically. For assumption three to hold true, these groups should focus more on formal traits and less on communicative functionality when using the NORS than when scoring impressionistically. The results presented in Table 47 and Table 48 reveal that this is not the case. Both percentages and frequencies show an increased focus on communicative functionality when using the NORS. The formal traits show only minor differences when focus is on the percentages of raters using them, but a considerable degree of decrease in use when frequency is considered. Both groups of SP-raters obtained higher estimates of IRR when scoring impressionistically, hence for assumption three to gain support for these groups, raters should focus less on formal correctness and more on communicatively related traits when using the NORS. This is indeed the case, and the results of the frequency table are particularly convincing.

On a whole, it seems that the three assumptions were more powerful as sources of explanations of rater reliability for group differences, that is in relation to H1, than for differences caused by scoring methods, in relation to H2.and particularly as an explanation to why naïve NS outperform all SP-raters in the NORS-based scoring, but not why the group of expert SP-raters outperforms the group of N2-teachers in the impressionistic scoring. In relation to H2, the assumptions are useful to some degree in explaining why the groups of all SP-raters and expert SP-raters obtain higher IRR estimates when scoring impressionistically, but not why the groups of naïve NS and N2-teachers score more reliably when using the NORS.

11.3 Discussion of the results of the qualitative study.

In this section I will discuss the main findings of the qualitative investigation. and try to explain some of the more unexpected results. I shall start with a discussion of the results produced by the study of the effect of rater training on criteria (presented in Section 10.1), touching on the results of the study regarding the effect of rating scale on criteria (presented in Section 10.2), and finally

I will discuss the results of the validity study (presented in Section 10.3). The main results of each study will be briefly repeated before each discussion.

11.3.1 A discussion of whether different rater groups focus on different criteria.

In the presentation of the results of the qualitative investigation, the first question posed was whether distinct rater groups focus on different aspects of speech in their ratings (Section 10.1). The discussion of this question is based on the results of the impressionistic scoring: the effect of rater training and experience is most evident when raters score without a rating scale, because, as we shall see, the use of the NORS eliminates group differences to a large extent. For the impressionistic scoring, the percentages and frequency calculation show similar results (Table 11 and Table 12).

- **There are no major differences between the groups in the use of the traits g*rammar, pronunciation* and *vocabulary,* which are used to a large extent by almost all raters of the different groups. The only case apart being the N2-teachers' lesser of focus on pronunciation**

Evidently, raters of all groups, naïve NS, N2-teachers and raters of Språkprøven consider formal correctness to play an important part in an L2-learner's oral performance. This is not very surprising*: grammar, pronunciation* and *vocabulary* are salient traits of performance and relatively easy to comment on as they refer to a norm of correctness. As native speakers, we have an intuition about what is correct and incorrect for these aspects even if we are not always capable of verbalising the rule that has been violated.

More interesting is the question why N2-teachers focus less on *pronunciation* than the other groups. One possible explanation may be that through their job, N2-teachers get so accustomed to Norwegian spoken with foreign accents, that they develop a high level of tolerance for these varieties. Probably, they are so used to listening to non-nativelike pronunciation that they are no longer disturbed by it.

While there are only minor differences between groups for the formal linguistic traits, there are rather considerable group differences for some of the other traits.

- **The focus on *communicative ability* correlates positively with rater training.**

This is as expected if trained raters internalise the criteria of the rating scale. *Communicative ability is* highlighted in the NORS as the primordial criteria. Since the naïve NS are unfamiliar with the scale, it is not surprising that they fail to focus on this trait. N2-teachers, however, focus more on

this trait than naïve NS, but less than the expert SP-raters. This may be due to the dominating position of communicative language teaching in Norway. Even without knowing the scale, through their education as language teachers, N2-teachers may have a concept of language ability as being something more than knowledge of its formal parts.

- **Rater training has a positive effect on raters' focus on a candidate's *comprehension*.**

The group of expert SP-raters is more focused than raters of the other groups on whether or not the candidate understands. This may once more reflect a positive effect of rater training: the NORS focuses on whether or not the interaction between candidate and examiner is meaningful and smooth. It is highlighted as an important aspect of oral performance to be able to transmit and understand information and avoid misunderstandings.

- **Expert SP-raters and naïve NS focus less on the trait *initiative* than the groups of N2-teachers and all SP-raters.**

This is indeed an unexpected finding. *Initiative* is one of the main criteria of the scale, yet expert SP-raters fail to use it in their ratings. There are no simple explanations to expert SP-raters' lack of focus on this trait when scoring impressionistically. If they had indeed internalised the criteria of the NORS as argued above, this should apply to *initiative* just as much as to *communicative ability*. The only explanation that comes to mind is that they may conceive this trait as an integrated aspect of communicative ability and therefore fail to mention it separately.

- **There is a positive effect of rater training on the use of the trait *strategic competence*.**

There is a clear positive correlation between rater training and the focus on *strategic competence* in raters' WR. No naïve NS refer to this aspect, a small percent of N2-teachers, close to 40 % of all the SP-raters and finally 67 % of the expert SP-raters do so. *Strategies* are not mentioned as a main criterion of the NORS, yet the scale refers to strategic competence in the level descriptors. *Strategic competence* is also referred to as part of the communicative competence framework upon which Språkprøven is based, and with which raters are made familiar through the rater training sessions.

- **L2 teachers are more focused on *content* than the other rater groups.**

Above, we saw that the group of N2-teachers focused less on *pronunciation* than the other rater groups. Another characteristic trait of this group, is its rather extensive focus on *content*. *Content* is not mentioned as a criterion of the NORS and only used by a small percentage of the expert SP-rater group. A few more raters of the group of all SP-raters focus on it, while almost no raters of the naïve NS group use this trait in their reports. What needs explaining here, then, is why almost half of the N2-teachers refer to content. Again the explanation is speculative, but I assume it may have to do with N2-teachers' habit of referring to a curriculum when assessing their pupils. Most tests used in an educational setting in Norway are achievement tests referring to what pupils or students have learnt in the course of instruction. Such tests often refer to a curriculum. Språkprøven, on the other hand, is a proficiency test and is not based on a curriculum[20]. The factual content of the performance of candidates should therefore not be taken into consideration. Without rater training, N2-teachers are not aware of this fact, and may easily continue to pay attention to the factual content of the candidates' performance.

In Chapter 10, the criteria were grouped in two main categories: formal linguistic traits and communicative functionality. This is perhaps where the effect of rater training is most clearly perceived. Again, I report the results of the impressionistic scoring as presented in Table 15 to Table 18. The main findings may be summarised as below.

- **There are only minor differences between the groups when it comes to their focus on *formal linguistic traits*. Naïve NS, however, focus more on *formal traits* than the other groups. For *communicative functionality*, on the other hand, there is a positive effect of rater training.**

There are no major group differences as to formal linguistic traits. It is, however, interesting that the group that focuses the most on these traits, is the group of naïve NS. These results are in line with those of Chalhoub-Deville (1995). She finds that non-linguist native speakers focused more on grammar-pronunciation than language teachers, contrary to the results of Galloway (1980) and Hadden (1991). One possible reason for the naïve NS' focus on formal traits, may be that these aspects refer to a norm of correctness of which native speakers have an intuition, as discussed earlier. These traits are therefore more salient to naïve NS than whether or not the message gets across. Another explanation may be that the naïve NS of the present study all have higher education. In Norway this means that they are most likely to know at least two foreign languages

[20] Even though it is used in establishing the proficiency level of adult second language learners in the state-run courses of Norwegian, and developped in accordance with Opplæringsplanen i norsk med samfunnsfag for voksne innvandrere 1998.

Being highly educated may also mean that they are used to analytic thinking, and possibly concerned by formal correctness. Correctness may very well be a condition in their professional contexts. The results would perhaps have been different if the naïve NS were less educated[21]. The main difference between the groups, is however not to be found in their reference to formal traits. All groups refer to these traits to a considerable degree, and these traits are also central in the NORS. Rather, the greatest differences between the groups are to be found in their focus on communicatively related traits. Of the four rater- groups, the group of naïve NS is the one that focuses the least on communicative functionality, N2-teachers use it a little more, yet less than the group of all SP-raters. The group of expert SP-raters uses it the most. Hence, for these traits there is a positive effect of rater training. These results match those of Halleck (1992) who found that trained raters are: "primarily concerned with communicative strategies rather than with the grammatical accuracy of the interviewees" (Halleck 1992:228). Through rater training and live ratings, SP-raters are made familiar with the NORS and the communicative framework upon which Språkprøven is based. They are taught to focus on formal linguistic traits, but always in relation to whether or not errors hinder communication. Even when scoring without the NORS at hand, SP-raters, and expert SP-raters in particular, manage to do this.

Other results presented in Chapter 10 in relation to group differences relate to the amount of criteria and the internal agreement between raters about the criteria:

- **The group of SP-raters uses a wider range of criteria in their WR than do the groups of informants without rater- training.**

Probably, rater training has resulted in making raters aware of a set of distinct aspects of speech. Informants without rater training, on the other hand, focus on a limited number of the most salient traits of speech, *grammar, vocabulary* and *pronunciation,* as we have seen. This is probably a natural consequence of all acquisition of knowledge: an ornithologist will probably notice a larger range of different birds than someone without that same knowledge of different bird species when going for a walk in the forest. Similarly, a trained rater will be capable of noticing and describing more aspects of N2-performance than a lay person.

21

In the study "Lekfolk og fagfolks vurdering av aksentpreget norsk" (Carlsen, In progress) ("Lay persons and professionals' assessment of Norwegian with an accent") the groups of informants of the present dissertation are compared to a group of lay persons without higher education in an attempt to check this assumtion.

- **Expert SP-raters are highly agreed about the criteria of speech. More surprisingly, so are the naïve NS.**

It is as expected that the expert SP-raters are agreed about the criteria of speech, one of the purposes of rater training being to create a common interpretation of the criteria and level descriptors of the rating scale, which White calls an "interpretation community" (White 1985). The internal agreement about criteria demonstrated by the group of expert SP-raters, is one sign of the effect of rater training. More surprisingly, the group of naïve NS also demonstrates a fairly high degree of internal agreement about the criteria. Where does this agreement come from? The informants of this group have different educational backgrounds: they are anthropologists, medical doctors, physicists, economists, lawyers and engineers etc. Their common characteristics are above all that they are native speakers of Norwegian on the one hand, and that they have higher education, on the other. Their common reference to formal traits may reflect native speakers' intuition about a norm of correctness, or it may reflect higher educated peoples' assumed focus on formal correctness, as argued above. The data do not lend themselves to an investigation of which of the two reasons is more likely to be true.

11.3.2 A discussion of whether raters focus on different criteria when they score impressionistically and NORS-based.

The second question posed in Chapter 10 was whether the use of the NORS affects the criteria raters use. In this section, differences from one scoring method to the other are focused on, while group differences are subordinate. In Table 19 percentages of all raters where grouped together for a study of the effect of rating scale on criteria. Table 20 presents the difference in frequency of use for all raters across scoring methods. The main effect of the NORS for all groups is presented and commented on below:

- **The percentage table shows that when all raters use the NORS *fluency* is referred to considerably less while *communicative ability* and *initiative* see a large increase in use. The other traits are used by a similar number of raters across scoring method. The frequency table shows similar results, in addition it reveals a considerable drop in the focus on *grammar.***

In other words, the focus of raters of all groups joined together is changed from *formal linguistic traits* and *fluency* when scoring impressionistically, to *communicative ability* and *initiative,* when scores are based on the NORS. The use of the scale is therefore able to change rater behaviour. It is understandable that the use of *communicative functionality* rise from impressionistic to NORS-based

scoring, since there is an explicit focus on communication skills in the scale. It is however hard to explain why the use of the NORS leads to less focus on *grammar,* this trait being one of the main criteria of the scale. One plausible explanation may be that the focus on whether or not the message is understandable, whether or not errors lead to misunderstandings and communicative breakdowns, undermine the role of grammatical correctness. Even though all candidates in this study commit grammatical errors, these do not always hinder communication.

Each group of informants was also studied separately for the effect of the rating scale on criteria, the main findings of the percentages and frequency tables (Table 23 and Table 24) were as follows:

- **The group of expert SP-raters is the one that varies the most from impressionistic to scale based scoring as to the criteria they use.**

This finding is particularly evident in the calculation of percentages of raters focusing on different traits, while the group differences are not so large for the frequency calculation. However, the finding that expert SP-raters vary the most from impressionistic to scale based scoring falsifies the assumptions that through rater training raters internalise the criteria and use them as a basis for their impressionistic scoring as well. This is surprising and I have no good explanations for these findings. The three criteria where expert SP-raters vary considerably more than the other groups, based on the percentages calculation presented in Table 23, are *grammar, intelligibility,* and *strategies*. The only possible reason I can find as to why trained and experienced raters use the criteria differently from impressionistic to NORS-based scoring is that they have automated the skill of setting a score according to the criteria of the NORS. The high inter-rater reliability estimates, together with their focus on communicative functionality when scoring impressionistically, could be a sign of this. When they are presented with the NORS and are asked to focus on the scale, raters may experience the same confusion as experienced typists when they start looking at the keyboard. The automated skill is disturbed by their focal attention, and they start committing errors in a skill that they normally perform without having to pay any attention to at all.

11.4 A discussion of the results of the validity investigation.

The main purpose of the qualitative data was as an empirical basis for testing H3 and H4, which postulate a positive effect of rater training and rating scale on construct validity, defined as the

match between raters' criteria and those of the NORS. We shall discuss the results in relating them to each hypothesis in turn, starting with the results of H3. The main findings were as follows:

- **H3, postulating a positive effect of rater training on construct validity, is partly sustained by the data. For most traits, there is a greater correlation between the criteria of the expert SP-raters and those of the scale than there is for the other rater groups.**

The group of expert raters focuses on communicative functionality to a larger extent than the other rater groups, without losing sight of the formal linguistic traits, which are also part of the NORS. Expert SP-raters manage to focus on formal traits in relation to whether or not these errors hinder communicative effect as emphasised in the rating scale. What is surprising, though, is the total lack of focus on the trait *initiative* when expert SP- raters score impressionistically. One would assume raters to emphasise this trait since it is given focal attention in the NORS, the principal criteria of the scale being communication and linguistic initiative. One explanation may be that experienced raters interpret initiative as part of the communicative ability and therefore fail to mention it as a separate trait.

It is worthy of comment that there are indeed differences between the group of all SP-raters and the group of expert SP-raters as to the criteria they use. As evident in Table 11 and Table 17, there is a poorer match between the criteria of the NORS and those of the group of all SP-raters than there is for the group of only expert SP-raters. This supports the assumption that it takes time to become a high-quality rater: the results of the present study show a positive effect of rater training on validity (as well as on reliability), but the effect is not immediate. Consequently, it is but after extensive rater training, consisting in rater training sessions as well as live ratings, that raters manage to conform to the rating scale and at the same time give reliable scores. This is further discussed in Chapter 12.

- **H4, which postulates a positive effect of rating scale on construct validity, is sustained by the data for all groups.**

When raters use the rating scale as a basis for their scorings, raters of all groups manage to a much greater degree to score in accordance with the underlying theoretical construct of Språkprøven as operationalised in the scale (Table 17). The focus on formal traits stays relatively constant, which is as expected since these traits are indeed part of the scale. However, there is a considerable increase in the use of the communicatively related traits in accordance with the highlighted position that these traits hold in the NORS. All groups focus more on

communicative aspects of speech when using the scale than when scoring impressionistically, yet the differences between the scoring methods are largest for the groups of naïve NS and L2 teachers, smaller for the group of all SP-raters and almost insignificant for the group of expert SP-raters. This is as expected, since the raters of Språkprøven focus on communicative functionality to a larger extent than the other groups even when scoring impressionistically. The traits *fluency* and *content,* which are used to some extent in the impressionistic scoring but which are not part of the NORS, see a considerable decrease in use when raters base their scores on the scale. An important finding of the qualitative study, then, is:

- **Even without rater training, the use of an explicit rating scale has a positive effect on construct validity**.

The use of the NORS is capable of changing rater behaviour from being one-sidedly focused on *formal correctness* to a greater emphasis on *communicative aspects* of speech, and thereby heightening the construct validity of a language test based on a model of communicative competence. The practical implications of this will be discussed in the next chapter.

CHAPTER 12: SUMMARY AND CONCLUSIVE REMARKS

The current research project has focused on the assessment of oral Norwegian as a second language in relation to a national N2-test, *Språkprøven i norsk for voksne innvandrere*. In order to ensure a fair assessment of the candidates' oral production, the test constructors of Språkprøven make use of *trained raters* basing their scores on an *explicit rating scale (NORS)*. These two highly recommended procedures in performance testing have traditionally been viewed as means to heighten reliability of test scores. In line with the recently developed concepts in this field, I argue that the rater variable affects not only reliability but the very construct validity of test scores: if raters fail to focus on the aspects of speech specified in the rating scale, this will affect construct validity just as if the rating scale itself were a poor operationalisation of the construct. Consequently, the use of a rating scale and trained raters are assumed to affect the construct validity as well as the reliability of test scores.

The development of a rating scale with precise level descriptors based on the underlying construct of the test, is a time consuming enterprise. So is the continuos training of raters. To establish the effect of these procedures is therefore relevant for theoretical as well as for practical and economic reasons. As a researcher, I am primarily interested in the theoretical implications of my study, yet test constructors and test users will probably be more interested in the practical implications. I will discuss both aspects in this chapter after a summary of the main results.

The project was guided by four main hypotheses, two of which concern the effect of rater training and rating scale on inter-rater reliability and two of which relate to their effect on construct validity. The first hypothesis:

- H1: **Training of raters** affects **reliability** of scores positively; trained raters show higher inter-rater reliability than untrained raters when scoring both with and without rating scales.

H1 was partly sustained by the data: the group of the most experienced raters showed the highest degree of internal agreement when scoring impressionistically, and was the only group that reached an acceptable level of reliability. However, when scores were based on the NORS, there were only insignificant differences between the groups, and the group that obtained the highest degree of reliability was the group of naïve NS, without any prior knowledge of the scale.

H2 addresses the effect of the use of rating scale on reliability:

- H2: The use of an explicit **rating scale (NORS)** affects **reliability** of scores positively; inter-rater reliability of scores is higher when raters use a rating scale (the NORS) as opposed to when they score impressionistically. The effect of a rating scale is positive for raters with and without rater training, yet the effect is greatest for the groups of untrained raters (naïve NS and N2-teachers).

H2 was supported by the data for the groups of informants without rater training (naïve NS and N2-teachers) but not for the two groups of SP-raters. The groups of naïve NS and N2-teachers show considerable increase of IRR estimates from impressionistic to NORS-based scoring method. All SP-raters show only minor differences across scoring methods, but they however obtain higher IRR estimates when scoring impressionistically than when using the NORS. The group of expert SP-raters shows considerably higher IRR when scoring impressionistically than when basing their scores on the NORS, contrary to the predictions of H2. As discussed in Chapter 11, there are no obvious reasons for this, except perhaps for the assumed confusion when focused attention is paid to a highly automated skill.

The next two hypotheses regard the effect of rater training and rating scale on construct validity defined for this project as the match between the criteria of the rating scale and those of the raters.

- H3: **Training of raters** affects **construct validity** (defined as the match between the criteria of the scale and those of the raters) positively: there is a greater match between the criteria of the NORS and those of the trained raters than between the NORS and the criteria used by other rater groups.

The qualitative data consist of both calculation of percentages of raters focusing on different traits, as well as calculations of the frequency by which each rater of different groups refer to the distinct criteria. More sophisticated statistical analysis would without doubt have been an advantage here. H3 did however gain support in the qualitative data for the most part the principal criterion of the NORS is *communication and linguistic initiative,* and the other criteria are the formal traits: *grammar, vocabulary* and *pronunciation.* There is a frequent reference in the scale descriptors of the NORS to strategies, comprehension and the ability to communicate a message in an understandable way. The formal traits are, on the whole, closely linked to whether or not errors hinder communication.

The group of expert SP-raters does, to a larger extent than the other rater groups, manage to uphold this linking of formal correctness to communicative functionality. In addition, expert SP-

raters focus more extensively on communicative functionality than the other groups (Table 17 and Table 18).

The next and last hypothesis postulates a positive effect of rating scale on construct validity:

- H4: The use of an explicit **rating scale (**NORS) affects **construct validity** (as defined in H3) positively. There is a greater match between the criteria of the NORS and those of the raters when raters base their scores on the NORS than when scoring impressionistically.

H4 does also gain support in the data: the formal linguistic traits stay relatively stable across scoring methods, the exception being less frequent reference to grammar when raters use the NORS. Communicative functionality, on the other hand, sees quite a considerable increase in use for all groups when raters use the scale. *Fluency* and *content*, the two criteria that are not mentioned in the NORS, but which were nevertheless used to a considerable extent in the impressionistic scoring, decline sharply in use when raters base their scores on the NORS. These results suggest that rating scale does indeed have a positive effect on construct validity even used in isolation without formal rater training.

The qualitative data were also used in an attempt to explain the results of the IRR study. Three assumptions were launched: IRR should be affected positively if:

- raters focus on a limited number of criteria
- there is high internal agreement about the criteria used
- raters focus on norm-referenced formal linguistic traits over communicative functionality which lacks reference to such a norm

When scoring impressionistically, the group of expert SP-raters obtained the highest degree of inter-rater reliability. The assumptions presented above were only in part suitable for explaining the agreement about scores of this group. Expert SP-raters did indeed show a higher degree of internal agreement about the criteria for speech than the other groups, as postulated by the second assumption. The other assumptions did not hold true, however: expert SP-raters focused on a larger set of criteria than the other groups, and they focused on communicatively related traits to a larger extent than the other groups when scoring impressionistically.

When scorings are based on the NORS, the group of naïve NS obtains the highest degree of inter-rater reliability (yet, we should not forget that there are minor differences between the groups for this scoring method). For this rater group, the three assumptions were indeed capable

of explaining how the group of naïve NS managed to obtain the highest estimates of reliability when using the NORS. They do so by simplifying the task of rating: they focus on few criteria, they are relatively in agreement about the criteria they use, and as compared to the other groups, they are primarily focused on norm-related formal traits and less on communicatively related aspects of speech.

12.1 Theoretical implications of the study

The present study supports the claim that rater training and the use of rating scale have a positive effect on test scores. It does also support the claim made in modern test research that the rater variable affects not only reliability but the construct validity of test scores since different rater groups focus on different aspects of speech. In order to ensure the validity of our test scores, we need to be aware of the underlying criteria of raters. I have shown that both rater training and the use of the NORS contribute to making raters score more in accordance with the underlying construct of the test as operationalised in the rating scale. An advantage of the present study is that rater training and the use of a rating scale are studied separately as well as together. This is a necessary design if one wishes to shed light on the effect of each procedure and the interaction between them. The most reliable scores are given by the group of expert raters of Språkprøven, yet the results show an increase of reliability for all raters when using the NORS. This means that even though the ideal would be to use a combination of trained raters and a rating scale, the use of rating scale does have a positive effect even when it is used by lay persons and N2-teachers without rater training.

As earlier quoted, Weigle claims that "[…] a de-emphasis on inter-rater agreement may have implications for the construct validity of the test if it draws attention away from getting raters to agree on a definition of the ability being measured by the test" (Weigle 1994:6). This linking of agreement about the scores and agreement about the underlying construct is supported in the present study. Indeed, the results suggest that agreement about the construct of the test is a prerequisite for agreement about test scores.

Another theoretical implication of this study, is that it shows the value of combining quantitative and qualitative approaches when investigating rater effect on test scores. Indeed, this study is consonant with Connor-Linton's claim that "quantitative similarities in ratings may mask significant qualitative differences in the reasons for those ratings" (Connor-Linton 1995). This is particularly evident in the ratings of the naïve NS. This group obtained the highest degree of IRR

estimates when scores are based on the NORS. Yet, when we looked behind the scores, we saw that this groups was more one-sidedly focused on formal correctness than the other rater groups. For this group, then, there was a negative correlation between IRR estimates and index of construct validity. This fact highlights the importance of complementing studies of rater reliability with studies of raters' underlying criteria, a point also made by Shohamy et al 1992 and Tarnanen 2002.

12.2 Practical implications of the study

In professional language testing, the use of trained raters and the development of explicit rating criteria and rating scales, are widely used procedures in assuring reliability of performance based tests. Modern test literature, this thesis being no exception, argues that these procedures also affect the construct validity of scores. In Norway these procedures are only used to a very limited degree, and the potential unreliability and lack of validity of untrained raters' scores, even at high-stake university exams, are rarely questioned (Berge 1993, Carlsen 2000, 2002). This is about to change with the introduction of the so-called reform of quality ("Kvalitetsreformen for høyere utdannelse") which is to be introduced in higher education from autumn 2003. The reform includes specifications of rating criteria and level descriptors for the distinct disciplines and subjects. This study is one argument in support of such an approach since for a test to yield reliable as well as construct valid scores it is a prerequisite that raters are trained and experienced in scoring according to an explicit rating scale which operationalises the underlying theoretical construct of the test.

The results are also in favour of the work done at Norsk språktest since the end of the 80s: the time- consuming and costly procedures of developing valid rating scales and training raters in interpreting and using these scales, do indeed bear fruits in that, not only does the group of expert SP-raters give more reliable scores, but, perhaps even more importantly, they manage to focus on the underlying construct of the test to a much greater degree than any other group.

However, two other results need a comment in relation to practical implications of the study: the group of trained raters of Språkprøven with varying degrees of experience, does not manage to give scores on an acceptable level of reliability. When scores are based on the NORS, naïve NS as well as N2-teachers without rater-training obtain higher reliability-coefficients. This may perhaps be a result of this group's more extensive focus on communicatively related traits when scoring with the rating scale than the two groups of informants without rater training (Table 17 and

Table 18). They try to conform to the construct, but by doing this they jeopardise reliability. This may be due to the difficulty of assessing the communicatively related traits in a reliable way. To do this successfully, extensive rater training, consisting in rater training sessions as well as live ratings, is necessary. The practical implication, then, is that the test constructors should keep up their work of rater training and rating scale development, but the degree of rater training needs to be heightened in order to assure reliable and valid scores. It is not until raters have taken part in about ten live ratings that they manage to agree both with the underlying construct of the test, as well as with each other about which scores to assign.

Another practical implication, which does to some extent contradict the above, is deduced by the support of H2 postulating a positive effect of rating scale on reliability. As we have seen, when raters based the scores on the NORS, group differences were close to eliminated. There were only insignificant differences between the groups for this scoring method. In addition, the group of naïve NS without rater training outperformed the groups of raters of Språkprøven. H4, which claimed a positive effect of the rating scale on construct validity, was also supported by the data: the use of an explicit rating scale, even without any kind of rater training, is able to enhance reliability as well as construct validity of test scores. These results are somewhat controversial, and their practical implications are not purely positive. I would strongly argue against using these results as an argument against rater training. On the contrary, I very much agree with Weigle that rater training is important in securing reliability as well as validity. The results do however imply that one may manage to some degree with the use of a rating scale alone. In cases where rater training for different reasons is impossible to undertake, one may increase reliability as well as construct validity of test scores by the use of common rating criteria and an explicit rating scale with clearly formulated level descriptors. These findings therefore have some important implications for large-scale testing. In Norway, language testing and scoring is not part of teachers' education, yet teachers develop and evaluate tests of language as well as for other school subjects. The implications of this study is that even though the ideal would be to give teachers rater training as well as rating scales for their assessment of pupils, they can to some degree manage with a rating scale alone.

12.3 Limitations of the study and call for further research

The results of this study should be interpreted with some care, partly because of the limitation of the data set and partly because of the simplicity of the statistical tools applied. The number of

informants taking part as raters is relatively large as compared to similar studies, yet the number of candidates is limited. The study should therefore be replicated with a larger set of candidates.

In addition it would have been interesting to approach the data with more sophisticated statistical tools. A fit-analysis should be conducted in order to eliminate mis-fitting raters. The reliability estimates could have been calculated by using different coefficients to see whether this would yield higher IRR-estimates for all groups. The qualitative data do indeed lend themselves to more advanced statistical methods, factor analysis or MFR. Even though the opportunity to conduct such analysis on the data was not present within the limitations of this project, the data are available for future research with other statistical tools.

The fact that raters of Språkprøven give one holistic score for each candidate complicates the task of investigating their underlying criteria. If they, on the other hand, had given separate scores for the distinct criteria, as they do in today's revised version of the test, it would have been easier to gain insight into the criteria upon which they base their scores (Norsk språktest 2003). In this study, the qualitative study is based on raters' reports of which traits they emphasise. We cannot really know for sure whether this is actually the case. I very clearly see the advantage of Shi's approach over my own where she asks raters not only to report the criteria for their scores, but in addition, to range their criteria according to their degree of importance for the total score.

Despite these limitations of data size and statistical sophistication of analysis, I would claim to have shed some light on the effect of rater training and rating scale on inter-rater reliability as well as on construct validity of test scores. The results have theoretical implications for the effect of the rater variable on test scores, and practical implications for the importance of rater training and the use of rating scale in order to assure a fair and valid measurement of oral second language performance.

REFERENCES

ACTFL. 1986. *ACTFL Proficiency Guidelines*. American Council on the Teaching of Foreign Languages, New York: Hastings-on-Hudson.

Alderson, J. C. 1991a. Bands and scores. In Alderson, J. C. and B. North (Eds.) 1991.

— 1991b. Giving students a sporting chance. Assessment by counting and by judging. In Alderson, J. C. and B. North (Eds.) 1991.

Alderson, J. C. and A. Hughes (Eds.) 1981. *Issues in Language Testing.* ELT Documents 111. London: The British Council.

Alderson, J. C. and B. North. (Eds.) 1991. *Language Testing in the 1990s: The Communicative Legacy.* London: Modern English Publications and the British Council.

Alderson, J. C. and J. Banerjee 2001: State-of-the-Art Review. Language testing and assessment (Part I). *Language Teaching* 34: 213-36.

— 2001: State-of-the-Art Review. Language testing and assessment (Part II). *Language Teaching* 35: 79-113.

American Psychological Association (APA). 1985. *Standards for Educational and Psychological Testing.* Washington, DC: American Psychological Association.

Andersen, R. O. 1995a. Mellomnivåtesten. Bakgrunn- begrunnelse- funksjon. *NOA* 18.

— 1995b. Hvorfor blir noen bedre i norsk? To målinger av språkferdigheter hos en gruppe voksne fremmedspråklige. In Thorseth, M. (Ed.) 1995.

— 1999. Å måle språkferdighet. In Hagen, J. E and K. Tenfjord (Eds.) 1999.

Angoff, W. H. 1971. Scales, norms and equivalent scores in Thorndike, R. L. 1971.

Anivan, S. (Ed.) 1991. *Current Developments in Language Testing.* Singapore: Regional Language Center.

Bachman, L. F. 1981. The construct validation of the FSI oral interview. *Language Learning* 31, 1: 67-86.

— 1990. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

— 1997 personal communication, e-mail.

Bachman, L. F. and A. Palmer. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 4: 449-65.

— 1996. *Language Testing in Practice*. Oxford: Oxford university press.

Bachman, L. F.; B. K. Lynch and M. Mason. 1995. Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing* 12, 2: 238-57.

Bachman, L. and A. Cohen (Eds.) 1998. *Interfaces Between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press.

Berge, K. L. 1993. Hvorfor er det så vanskelig å få til en pålitelig bedømming av elevtekster? Om tekstevalueringens kroniske elendighet illustrert med eksempler fra sensuren i norsk hovedmål våren 1992. In Michelsen, P. A. (Ed.) 1993.

— 1996. *Norsksensorenes tekstnormer og doxa. En kultursemiotisk og sosiotekstologisk analyse.* Ph.D.dissertation. Department of applied linguistics, NTNU, Trondhjem.

Bloomfield, L. 1914. *An Introduction to the Study of Language*. New York: Henry Holt and Company.

— 1933. *Language*. New York: Holt, Rinehart and Winston.

Brindley, G. 1991. Defining language ability: The criteria for criteria. In Anivan, S. (Ed.) 1991.

— 1998. Describing language development? Rating scales and SLA. In Bachman, L.

and A. Cohen (Eds.) 1998.
Brodersen, R. and T. Kinn (Eds.) 2000. *Språkvitskap og vitskapsteori. Ti nye vitskapsteoretiske innlegg.* Larvik: Ariadne Forlag.
Brown, A. 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing* 12, 1: 1-15.
Brown, G. and G. Yule 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.
Brumfit, C. J. and K. Johnson (Eds.) 1979. *The Communicative Approach to Language Teaching.*
Oxford: Oxford University Press.
Buck, K. (Ed.) 1989. *The ACTFL oral proficiency interview. Tester training manual.* New York: The American Council on the Teaching of Foreign Languages.
Burt, M.; H. Dulay and M. Finocchiaro (Eds.). 1977. *Viewpoints on English as a Second Language.* New York: Regents.
Butler, C. 1985. *Statistics in Linguistics.* Oxford: Basil Blackwell Ltd.
Bygate, M. 1987. *Speaking.* Oxford: Oxford University Press.
Byrnes, H. 1989. The rating scale. In Buck, K. (Ed.) 1989.
Campbell, D. T. and D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56: 81-105.
Canale M. and M. Swain. 1979. *Communicative Approaches to Second Language Teaching and Testing.* Ontario: The Ontario Institute for Studies in Education.
— 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1: 1-47.
Canale, M. 1983: From communicative competence to communicative language pedagogy. In Richards, J. C. and R. W. Schmidt. (Eds.) 1983.
Carlsen, C. 2000. Den objektive språktesten: mål eller minne? In Brodersen, R. and T. Kinn (Eds.) 2000.
— 2002 Sensur og kvalitet. Feature article in Bergens Tidende. Mai 15th 2002.
— 2003. Et forsøk på å forklare sensorenighet in W. Vagle. 2003.
— (In progress). Fagfolk og lekfolks vurdering av aksentpreget norsk. Den nye norsken. Bergen: Universitetet i Bergen.
Carroll, J. B. 1961: Fundamental considerations in testing for English language proficiency of foreign students in *Testing the English Proficiency of Foreign Students* 31-40. Washington: Center for Applied Linguistics.
— 1968: The psychology of language testing. In Davies, A. (Ed.). 1968.
Chalhoub-Deville, M. 1995. Deriving oral assessment scales across different tests and rater groups. *Language Testing* 12, 1: 16-34.
Chomsky, N. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass: MIT Press.
— 1976. *Reflections on Language*. London: Temple Smith.
— 1981. *Lectures on Government and Binding.* Dordrecht: Foris.
— 1986a. *Knowledge of Language. Its Nature, Origin and Use.* New York: Praeger.
— 1986b. *Barriers.* Cambridge, Mass: MIT Press.
Clapham, C. and D. Corson. (Eds.) 1997. *Encyclopedia of Language and Education, Vol. 7, Language Testing and Assessment.* Dordrecht: Kluwer Academic Publishers.
Common European Framework of Reference for Languages: Learning, teaching, assessment. 2001.Cambridge: Cambridge University Press.
Connor-Linton, J. 1995. Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes* 14, 1: 99-115.
Cooper, C. R. 1977. Holistic evaluation of writing. In Cooper, C. R. and L. Odell (Eds.) *Evaluating writing: Describing, measuring, judging.* Urbana, IL: NCTE.

Crocker, L. and J. Algina. 1986. *Introduction to Classical & Modern Test Theory.* New York: Harcourt Brace Jovanovich College Publishers.
Crystal, D. 1997. *Dictionary of Linguistics and Phonetics*. Oxford: Blackwell Publishers.
Cumming, A. 1990. Expertise in evaluating second language compositions. *Language Testing* 7, 1: 31-51.
— 1997. The Testing of Writing in a Second Language. In Clapham, C. and D. Corson (Eds.) 1997.
Cumming, A. and R. Berwick (Eds.) 1995. *Validation in Language Testing*. Clevedon: Multilingual Matters Ltd.
Davies, A (Ed.). 1968. *Language Testing Symposium. A Psycholinguistic Perspective.* London: Oxford University Press.
Dechert, H. 1983. How a story is done in a second language in Færch, C. and G. Kasper (Eds.) 1983.
de Jong, J. 1988. Rating scales and listening comprehension. *Australian Review of Applied Linguistics.* 11, 2: 73-87.
Dregelid, K. M. Samtaleferdigheter i norsk som andrespråk på et mellomnivå*. Nordica Bergensia 26: 221-35.* Bergen: Nordisk institutt, Universitetet i Bergen.
Edgeworth, F. Y. 1890. The element of chance in competitive examinations. *Journal of the Royal Statistical Society* 53: 644-63.
Ellis, R. 1994. *The Study of Second Language Acquisition*. Oxford: Oxford University Press.
Evensen, L. S. and F. Hertzberg. 2001. "Læringsutbyttet i norsk skriftlig i grunnskolen: KAL-prosjektet" in Kulbrandstad, L. I. and G. Sjølie (Eds.) 2001.
Faarlund, J. T.; S. Lie, and K. I. Vannebo. *Norsk referansegrammatikk.* Oslo: Universitetsforlaget.
Fulcher, G. 1993. *The Construct Validation of Rating Scales for Oral Tests in English as a Foreign Language.* PhD thesis. Lancaster: University of Lancaster.
— 1996a. Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing* 13, 208-38.
— 1996b. Testing tasks: issues in task design and the group oral. *Language Testing* 13,1:23-51.
— 1997. The testing of speaking in a second language. In Clapham, C. and D. Corson (Eds.)1997.
Færch, C. and G. Kasper (Eds.) 1983. *Strategies in Interlanguage Communication.* London: Longman.
Galloway, V. B. 1980. Perceptions of the communicative efforts American students of Spanish. *Modern Language Journal* 64, 428-33.
Gingras, R. (Ed.) 1978. *Second Language Acquisition and Foreign Language Teaching.* Washington: Center for Applied Linguistics.
Grice, P. 1989. *Studies in the Ways of Words*. Cambridge: Harvard University Press.
Hadden, B. 1991. Teacher and nonteacher perceptions of second-language communication. *Language Learning* 41, 1: 1-24.
Hagen, J. E. *Norsk grammatikk for andrespråkslærere*. Oslo: Ad Notam Gyldendal.
Hagen, J. E. and K. Tenfjord. 1999. *Andrespråksundervisning*. Oslo: Ad Notam Gyldendal.
Hagen, J. E.; R. O. Andersen, M. H. Kløve, A. Kristiansen and K. Tenfjord (Eds.) 2003. *I Mannes Minne: Minneskrift over Gerd Manne.* Oslo: Fag og Kultur.
Halleck, G. 1992. The oral proficiency interview. Discrete Point test or a measure of communicative language ability. *Foreign Language Annals* 25, 3: 227-31.
— 1995. Assessing oral proficiency: a comparison of holistic and objective measures.

The Modern Language Journal 79:223-34.
Halliday, M. A. 1973: Relevant models of language. In M. A. Halliday 1973.
— 1973. *Explorations in the Functions of Language. London: Arnold.*
— 1976. The form of a functional grammar. In Kress, G. (Ed.) 1976.
Halliday, M. A.; A. McIntosh and P. Strevens. 1964. *The Linguistic Science and Language Teaching.* Bloomington: Indiana University Press.
Halvorsen, B. 2002. Uttale av norsk som andrespråk på et mellomnivå. *Nordica Bergensia 26:*163-81. Bergen: Nordisk institutt, Universitetet i Bergen.
— 2003. Vurdering av muntlig språkferdighet. In Vagle, W. 2003.
Hamp-Lyons, L. 1991a. Scoring Procedures for ESL Contexts. In Hamp-Lyons, L.(Ed.) 1991b.
— (Ed.) 1991b. *Assessing Second Language Writing in Academic Contexts.* Norwood: Ablex Publishing Corporation.
Harley, B.;P. Allen, J. Cummins and M. Swain (Eds.) 1990. *The Development of Second Language Proficiency.* Cambridge: Cambridge University Press.
Harris, D. P. 1968. *Testing English as a Second Language.* New York: McGraw Hill.
Harris, R. A. 1993. *The Linguistics Wars.* London: Oxford University Press.
Hasselgren, A. 1998. *Smallwords and Valid Testing.* PhD dissertation. Bergen: Department of English, University of Bergen.
Hatch, E. 1978. Discourse analysis and second language acquisition. In Hatch, E. (Ed.) 1978
— 1978. (Ed.): *Second Language Acquisition: A Book of Readings*: Rowley: Newbury House.
Hatch, E. and A. Lazaraton. 1991. *The Research Manual. Design and Statistics for Applied Linguistics*. Boston, Mass.: Heinle & Heinle Publishers.
Hellekjær, G. O. (In Progress) *The Acid Test: From Upper-Secondary EFL Instruction to the Reading of English Textbooks at Norwegian Colleges and Universities*. Oslo: Department of Education and School Development, University of Oslo.
Higgs, T. V. (Ed.) 1984. *Teaching for proficiency: The organizing principle.* Lincolwood, IL: NationalTextbook Co.
Hill, K. 1996. Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In Huhta, A.; V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.) 1997.
Hughes, A. 1989. *Testing for Language Teachers.* Cambridge: Cambridge University Press.
Huhta, A.; V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.)1997. *Current Developments and Alternatives in Language Assessment-Proceedings of LTRC 96.* Jyväskylä: University of Jyväskylä .
Huot, B. 1990. Reliability, validity and holistic scoring: what we know and what we need to know. *College Composition and Communication* 41: 201-13.
Hymes, D. 1972. On communicative competence. In Pride, J. B. and J. Holmes (Eds.) 1972.
— 1979. On communicative competence. Extensive extracts of the 1972 article. In Brumfit, C. J. and K. Johnson (Eds) 1979.
Joos, M. 1967. *The Five Clocks.* New York: Harcourt Brace Javanovich.
Kaftandjieva, F. and S. Takala. 2003. *Development and Validation of Scales of Language Proficiency.* In Vagle, W. 2003.
KAL-prosjektet. **http://prosjekt.hihm.no/r97-kal** or Evensen, L. 2001.
Kenyon, D. 1992. Introductory remarks at symposium on *Development and use of rating scales in language testing,* 14th Language Testing Research Colloquium. Vancouver,

February 27th –March 1st.
Krashen, S. 1977. The Monitor Model for second language performance. In Burt, M.; H. Dulay and M. Finocchiaro (Eds.) 1977
— 1978. Adult second language acquisition and learning: a review of theory and practice In Gingras, R. (Ed.) 1978.
Kulbrandstad, L. I. and G. Sjølie (Eds.) 2001. *På Hamar med norsk. Rapport fra konferansen "Norsk på mellomtrinnet",* 18.-19. januar 2001. Del I: Skriving og lesing. Høgskolen i Hedmark rapport nr. 11.
Lado, R. 1961. *Language testing. The Construction and Use of Foreign Language Tests*. New York: McGraw-Hill.
Linacre, J. M. 1989. Many-faceted Rasch measurement. MESA Press, Chicago IL.
— 1999. Measurement of Judgements. In Masters, G. N. and J. P Keeves (Eds.) 1999.
Linn, R. L .1989. *Educational Measurement.* New York: American Council on education.
Liskin-Gasparro, J. 1984. The ACTFL guidelines. A historical perspective. In Higgs, T. V. (Ed.) 1984.
Lumley, T. (In progress) Abstract of *The process of the assessment of writing performance: the rater's perspective.* PhD dissertation. Melbourne: Department of Linguistics and Applied Linguistics, The University of Melbourne.
Lumley, T. and T. McNamara. 1995. Rater characteristics and rater bias: implications for training. *Language Testing* 12, 1: 54-71.
Lynch, B. and T. McNamara. 1998. Using G-theory and Many-facet Rasch measurement in the development of performance assessment of the ESL speaking skills of immigrants. *Language Testing* 15,2:158-80.
Masters, G. N. and J. P. Keeves (Eds.) 1999. *Advances in measurement in educational research and assessment*. Amsterdam: Pergamon.
McLaughlin, B. 1987. *Theories of Second-Language Learning.* London: Edward Arnold
McNamara, T. 1996. *Measuring Second Language Performance.* London: Longman.
— 1997. Performance Testing. In Clapham, C. and D. Corson (Eds.) 1997.
— T. 2000: *Language Testing.* Oxford: Oxford University Press.
Meisel, J. M.; H. Clashen and M. Pienemann.1981. On determining developmental stages in natural second language acquisition. *Studies in Second Language Acquisition* 3: 109-35.
Messick, S. A. 1975. The standard problem: meaning and values in measurement and evaluation. *American Psychologist* 30: 955-66.
— 1989. Validity. In R. L. Linn1989.
— 1996. Validity and washback in language testing. *Language Testing* 13, 3: 241-56.
Michelsen, P. A. 1993. *Sjangeroppbrudd – Om stilskriving og skriveopplæring i den videregående skolen.* Oslo: Cappelen Forlag.
Milanovic, M.; N. Saville and S. Shuong. 1996. A study of the decision-making behaviour of composition markers. In Milanovic, M. and N. Saville (Eds.) 1996.
Milanovic, M.; N. Saville, A. Pollitt and A. Cook. 1995. Developing rating scales for CASE: Theoretical concerns and analyses. In Cumming, A. and R. Berwick (Eds.) 1995.
Milanovic, M. and N. Saville (Eds.) 1996. *Performance Testing, Cognition and Assessment. Selected papers form the 15th Language Testing Research Colloquium, Cambridge and Arnhem*. Cambridge: Cambridge University Press.
Moe, E. 2003a. Den gode prøven- finst han? In Vagle, W. 2003.
— 2003b. Eigenvurdering av språklege ferdigheiter- mål eller middel? In Hagen, J. E.; R. O. Andersen, M. H. Kløve, A. Kristiansen and K. Tenfjord 2003.
Moe, E. and N. Jones 2003. Using Multi-faceted Rasch analysis to validate a test of

writing. *Acta Didactica* 1.
Morrow, K. 1979. Communicative language testing: revolution or evolution? In Brumfit, C. J. and K. Johnson (Eds.) 1979.
Nickel, G. (Ed.) 1976. *Proceedings of the Fourth International Congress of Applied Linguistics*. Stuttgart: Hochschulverlag.
Niedzielski, N. A. and D. R. Preston. 2000. *Folk linguistics.* Berlin: de Gruyter.
Norsk språktest.1998a. *ALTES terminologiliste for språktesting, norsk–engelsk versjon.* Bergen: Universitet i Bergen.
— 1998b. *Språkprøven i norsk for fremmedspråklige voksne. Eksaminator og sensormappe.* Bergen: Universitetet i Bergen og Folkeuniversitetet, Norsk språktest.
— 2003. *Språkprøven i norsk for voksne innvandrere. Muntlig språkbruk*. Bergen: Universitetet i Bergen og Folkeuniversitetet, Norsk språktest.
Odlin, T. 1989 *Language Transfer. Cross –linguistic influence in language learning.* Cambridge: Cambridge University Press.
Oller, J. W. Jr. (Ed.) 1983. *Issues in language testing research.* Rowley: Mass.: Newbury House.
Pawley, A. and F. H. Syder 1983. Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In Richards J. C. and R. W. Schmidt (Eds.) 1983.
Pedersen, J. 1997. *Kva måler Språkprøven? ei sving i ein valideringsspiral.* Hovedfagsavhandling. Bergen: Nordisk institutt, Universitetet i Bergen.
Pienemann, M. 1998. *Language Processing and Second Language Development. Processability Theory.* Amsterdam: Benjamins.
Pilliner, A. E. 1968. Subjective and objective testing. In Davies, A. (Ed.) 1968.
Pollitt, A. and N. Murray. 1996. What raters *really* pay attention to. In Milanovich, M. and N. Saville (Eds.) 1996.
Popper, K. R. 1959. *The Logic of Scientific Discovery.* London: Hutchinson of London.
Pride, J. B and J. Holmes (Eds.) 1972. *Sociolinguistics*. Harmondsworth: Penguin Books.
Raaheim, A. 2000. Inter bedømmer reliabilitet ved eksamen på psykologi grunnfag. *Tidsskrift for Norsk Psykologiforening* 37: 203-13.
Raaheim, A. og K. Raaheim (Eds.) 2002. *Eksamen-en akademisk hodepine : en håndbok for studenter og lærere.* Bergen: Sigma forlag.
Richards, J.; J. Platt and H. Platt. 1992. *Longman Dictionary of Language Teaching & Applied Linguistics.* Essex: Longman Group.
Richards, J. C. and R. W. Schmidt. (Eds.) 1983. *Language and communication.* London: Longman.
Roeper, T. and E. Williams (Eds.) 1987. *Parameter Setting.* Dordrecht: D. Reidel Publishing Company.
Saleva, M. 1997. *Now They're Talking. Testing Oral Proficiency in a Language Laboratory*. PhD Thesis. Jyväskylä: University of Jyväskylä.
Salomonsen, L. E. 2002. Substantivfraser i norsk som andrespråk på et mellomnivå. *Nordica Bergensia* 26: 211-220. Nordisk institutt, Universitetet i Bergen.
Sapir, E. 1949 [1921] *Language. An Introduction to the Study of Speech*. New York: Harcourt Brace.
Savignon, S. 1997: *Communicative Competence. Theory and Classroom Practice*. New York: The McGraw-Hill Companies.
Searle, J. R. 1969. *Speech acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.
Saville, N. and P. Hargreaves. 1999. Assessing speaking in the revisited FCE. *ELT Journal* 53,1:42-51.
Schoonen, R.; M. Vergeer and M. Eiting. 1997. The assessment of writing ability: expert

readers versus lay readers. *Language Testing* 14, 2: 157-84.
Shi, L. 2001.Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing* 18, 3: 303-25.
Shohamy, E. 1981. The stability of oral proficiency assessment on the oral interview testing procedure. *Language Learning* 33, 4: 527-40.
— 1983. Interrater and intrarater reliabiltiy of the oral interview and concurrent validity with cloze procedure in Hebrew. In Oller, J. W. Jr. (Ed.) 1983.
Shohamy, E.; C. Gordon and R. Kraemer 1992. The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal* 76: 27-33.
Sieloff Magnan, S. 1987. Rater reliability of the ACTFL oral proficiency interview. *The Canadian Modern Language Review 43,2: 525-37.*
Skinner, B. F. 1957. *Verbal behavior.* New York: Appleton-Century-Crofts.
Spolsky, B. 1976. Language testing: art or science? In Nickel, G. (Ed.) 1976.
— 1985. The limits of authenticity in language testing. *Language Testing* 2, 1: 31-40.
— 1995. *Measured Words.* London: Oxford University Press.
Stenström, A. B. 1994. *An Introduction to Spoken Interaction.* London: Longman.
Tannen, D. 1984. The pragmatics of cross- cultural communication. *Applied Linguistics* 5, 3: 189-95.
Tarnanen, M. 2002. Arvioija valokeilassa: suomi toisena kielenä kirjoittamisen arviointia. [The rater in a spotlight: Rating Finnish as a second language writing]. Jyväskylä: University of Jyväskylä.
Thorndike, R. L 1971. *Educational Measurement.* Washington, DC: American Council on Education.
Thorseth, M. (Ed.) 1995. *Norskopplæring som virkemiddel i integreringsarbeidet?: En evaluering av dagens kvalifiseringsordninger for fremmedspråklige*. Trondheim: SINTEF rapport.
Underhill, N. 1987. *Testing Spoken Language. A handbook of oral testing techniques.* Cambridge: Cambridge University Press.
Upshur, J. and C. Turner. 1999. Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* 16,1:82-111.
Vagle, W. (Ed.) 2003: *Vurdering av språkferdighet.* Trondheim: Institutt for språk og kommunikasjon, NTNU.
Vaughan, C. 1991. Holistic assessment: What goes on in the rater's mind? In Hamp-Lyons, L. (Ed.) 1991.
Vygotsky, L. S. 1978. *Mind in Society. The Development of Higher Psychological Processes*. Cambridge Mass.: Harvard University Press.
Weigle, S. C. 1994. *Effects of Training on Raters of English as a Second Language Compositions: Quantitative and Qualitative Approaches.* PhD dissertation. Los Angeles: University of California.
— 1998. Using FACETS to model rater training effects. *Language Testing* 15 ,2: 263-87.
Weir, C. 1990. C*ommunicative Language Testing.* New York: Prentice Hall.
Weiss, C. H. 1972. *Evaluation Research: Methods for Assessing Program Effectiveness.* Englewood Cliffs, New York: Prentice Hall.
White, E. M. 1985. *Teaching and assessing writing.* San Fransisco: Jossey-Bass Publishers.
White, L. 1989. *Universal Grammar and Second Language Acquisition.* Amsterdam: John Benjamins Publishing Company.
Widdowson, H. G. 1983. *Learning Purpose and Language Use.* London: Oxford University Press.

Wigglesworth, G. 1993. Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing* 10,3:305-35.
Williams, E. 1987. Introduction. In Roeper, T. and E. Williams (Eds.) 1987.